AD_____

Award Number: DAMD17-98-2-8005

TITLE: Malaria Genome Sequencing Project

PRINCIPAL INVESTIGATOR: Malcolm J. Gardner, Ph.D.

CONTRACTING ORGANIZATION: The Institute for Genomic Research
Rockville, Maryland 20850

REPORT DATE: January 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2000	Annual (17 Dec 98 - 16 Dec 99)

4. TITLE AND SUBTITLE
Malaria Genome Sequencing Project

5. FUNDING NUMBERS
DAMD17-98-2-8005

6. AUTHOR(S)
Malcolm J. Gardner, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Institute for Genomic Research
Rockville, Maryland 20850

E-MAIL:
Gardner@tigr.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. To date, we have published the first complete sequence of a malarial chromosome (chromosome 2 [4]), completed the random phase sequencing of 3 other large chromosomes totaling 7.2 Mb, and have initiated functional genomics studies using glass slide micorarrays to characterize the expression of chromosome 2, 3, and 14 genes throughout the erythrocytic cycle. We have also collaborated in the construction of a two-enzyme optical restriction map of the entire *P. falciparum* genome [7] , and are continuing to further develop the GlimmerM gene finding software developed in year 1. In addition, we have begun small scale sequencing of the rodent malaria *P. yoelii* and are collaborating in the sequencing of part of a *P. vivax* chromosome. Discussions with the Malaria Program, NMRC aimed at development of a program to use genomics and functional genomics to accelerate vaccine research are in progress.

14. SUBJECT TERMS
Plasmodium falciparum, malaria, genome, chromosome, sequencing, Microarray, software

15. NUMBER OF PAGES
53

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

N/A Where copyrighted material is quoted, permission has been obtained to use such material.

N/A Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

MTB Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

N/A For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

PI - Signature Date

Table of Contents

Introduction

Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. It is a major public health problem in many tropical areas of the world, and also affects many individuals and military forces that visit these areas. In 1994 the World Health Organization estimated that there were 300-500 million cases and up to 2.7 million deaths caused by malaria each year, and because of increased parasite resistance to chloroquine and other antimalarials the situation is expected to worsen considerably [1]. These dire facts have stimulated efforts to develop an international, coordinated strategy for malaria research and control [2]. Development of new drugs and vaccines against malaria will undoubtedly be an important factor in control of the disease. However, despite recent progress, drug and vaccine development has been a slow and difficult process, hampered by the complex life cycle of the parasite, a limited number of drug and vaccine targets, and our incomplete understanding of parasite biology and host-parasite interactions.

The advent of microbial genomics, i.e. the ability to sequence and study the entire genomes of microbes, should accelerate the process of drug and vaccine development for microbial pathogens. As pointed out by Bloom, the complete genome sequence provides the "sequence of every virulence determinant, every protein antigen, and every drug target" in an organism [3], and establishes an excellent starting point for this process. In 1995, an international consortium including the National Institutes of Health, the Wellcome Trust, the Burroughs Wellcome Fund, and the US Department of Defense was formed (Malaria Genome Sequencing Project) to finance and coordinate genome sequencing of the human malaria parasite *Plasmodium falciparum*, and later, a second, yet to be determined, species of *Plasmodium*. Another major goal of the consortium was to foster close collaboration between members of the consortium and other agencies such as the World Health Organization, so that the knowledge generated by the Project could be rapidly applied to basic research and antimalarial drug and vaccine development programs worldwide.

Body

This report describes progress in the Malaria Genome Sequencing Project achieved by The Institute for Genomic Research and the Malaria Program, Naval Medical Research Center, under Cooperative Research Agreement DAMD17-98-2-8005, over the 12 month period from Dec. '98 to Dec '99. The specific aims of the work covered under this cooperative agreement were to:

1. Determine the sequence of 3.5 megabases of the *P. falciparum* genome (clone 3D7):

a) Construct small-insert shotgun libraries (1-2 kb inserts) of chromosomal DNA isolated from preparative pulsed-field gels.

b) Sequence a sufficiently large number of randomly selected clones from a shotgun library to provide 10-fold coverage of the selected chromosome.

c) Construct P1 artificial chromosome (PAC) libraries (inserts up to 20 kb) of chromosomal DNA isolated from preparative pulsed-field gels.

d) If necessary, generate additional STS markers for the chromosome by i) mapping unique-sequence contigs derived from assembly of the random sequences to chromosome, ii) mapping end-sequences from chromosome-specific PAC clones to YACs.

e) Use TIGR Assembler to assemble random sequence fragments, and order contigs by comparison to the STS markers on each chromosome.

f) Close any remaining gaps in the chromosome sequence by PCR and primer-walking using *P. falciparum* genomic DNA or the YAC, BAC, or PAC clones from each chromosome as templates.

2. Analyze and annotate the genome sequence:

a) employ a variety of computer techniques to predict gene structures and relate them to known proteins by similarity searches against databases; identify untranslated features such as tRNA genes, rRNA genes, insertion sequences and repetitive elements; determine potential regulatory sequences and ribosome binding sites; use these data to identify metabolic pathways in *P. falciparum.*

3. Establish a publicly-accessible *P. falciparum* genome database and submit sequences to GenBank.

We are pleased to report that excellent progress has been made towards achievement of these goals. In last year's annual report we announced the the publication in *Science* of the first complete sequence of a malarial chromosome (chromosome 2) [4]. In addition, we reported on work done by the TIGR/NMRC team and collaborators to provide new tools and resources for the Malaria Genome Project, including development of a *Plasmodium* gene finding program, GlimmerM [5], and introduction of optical restriction mapping technology for rapid mapping of whole *Plasmodium* chromosomes [6]. We also reported that sequencing of 3 additional *P. falciparum* chromosomes was underway, and that we were investigating the use of microarray technology to examine the expression of all genes from chromosomes 2 and 3 of *Plasmodium.* To facilitate community access to the sequence data, a *P. falciparum* genome web site was also established at TIGR which contains all of the chromosome 2 sequence data and annotation, as well as preliminary data for other chromosomes currently being sequenced (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html).

In the past year we have completed the high-throughput sequencing phase of chromosomes 10, 11, and 14, which together account for 30 % of the genome. These chromosomes are now in the gap closure phase, and chromosome 14 is expected to be completed this year, and chromosomes 10 and 11 will be completed shortly after. We also collaborated with David Schwartz's laboratory in construction of a two-enzyme optical restriction map of the entire *P. falciparum* genome; this was published recently in Nature Genetics [7]. As indicated in

last year's report we also initiated a functional genomics program in collaboration with the Malaria Program, NMRC. Glass slide microarrays containing PCR fragments from almost all genes from chromosomes 2 and 3 have been prepared, and experiments to profile the expression of these genes through the erythrocytic stage of the life cycle are underway. We have also assisted NMRC in their pilot project to apply the techniques of proteomics towards the identification of novel antigens in parasite (sporozoite) extracts. Finally, we are currently reviewing with NMRC further steps that can be taken to more rapidly apply *Plasmodium* genomics, functional genomics, and proteomics to problems of vaccine development for malaria.

Sequencing of *P. falciparum* chromosome 14 (Specific Aim 1)

Sequencing of chromosome 14 (3.4 Mb) is being funded primarily by a grant from the Burroughs Wellcome Fund; funds from this collaborative agreement are being used to accelerate the sequencing, assist in closure and annotation, develop microarrays for chromosome 14, and facilitate rapid utilization of the sequence data by the DoD vaccine and drug development groups. In last year's report we described the isolation of chromosome 14 DNA on pulsed field gels, preparation of shotgun libraries in pUC18, and high throughput sequencing of these libraries. The high-throughput sequencing phase of the project was completed in December 1998. 74,292 sequences with an average read length of 530 nt were produced. All of these sequences were performed with FS+ dye terminator chemistry which we had previously found to be superior to dye primer chemistry for the sequencing of AT-rich *P. falciparum* DNA. This is equivalent to 9X coverage assuming that due to co-migration of sheared nuclear DNA with the chromosome 14 DNA on pulsed field gels, 20% of sequences in the shotgun library were derived from other chromosomes. The sequences were assembled in a 2 step procedure with TIGR Assembler [8]. The first assembly was performed at 99.5% stringency to produce a robust set of contigs; these contigs and the remaining unassembled sequences were then used to start a second assembly at 97.5% stringency. 1,750 contigs were obtained and the largest contig was 99 kb. In comparison, the largest contig obtained after the first assembly of the chromosome 2 data was about 20 kb, indicating that exclusive use of the dye terminator chemistry for chromosome 14 resulted in the production of high quality sequence data.

The gap closure process began in December 1998. The procedures being used to close gaps are basically the same as those used previously on the chromosome 2 project[4], namely 1) use of GROUPER software to identify groups (contigs linked by shotgun clones), physical gaps and sequence gaps; 2) editing of contigs ends to remove untrimmed vector sequence, low quality sequence data, and chimeric clones that prevent merging of contigs; 3) resequencing of missing mates and short sequences at contig ends; 4) sequencing of shotgun clones spanning sequence gaps using primers at the ends of the gaps; 5) PCR with genomic DNA to span physical gaps; and 6) use of the transposon insertion method to close very AT-rich gaps. In practice, GROUPER is run on the set of contigs produced by an assembly and some or all of steps 1-6 are performed until no further progress is possible. Another assembly is then performed with the edited contigs, new sequences (*e.g.* primers walks and missing mates), and unassembled sequences left over from the previous assembly. The new assembly will incorporate new sequences such as primer walks produced during closure, sequences edited during closure, and other sequences that did not get merged into the previous assembly, thereby providing new starting points for additional work. This process is repeated until the sequence is closed.

As noted above, due to cross-contamination of the chromosome 14 DNA with sheared nuclear DNA, up to 20% of the sequence data is derived from chromosomes other than chromosome 14. In order to focus the closure efforts on chromosome 14 contigs, chromosome 14 markers are used to identify which contigs and groups of contigs are from chromosome 14. With chromosome 2 about 30 markers were available (1 marker per 30 kb). In contrast, for chromosome 14 there are 98 STS markers derived from YACs (provided by Alister Craig) plus an additional 101 SSLP markers [9], providing a marker about every 17-20 kb. The higher density of markers will allow identification of more chromosome 14 contigs and should simplify the gap closure process. In addition, with funding provided by the BWF, David Schwartz's group has completed a 2-enzyme optical restriction map of the *P. falciparum* genome [7]. We will use the optical map and the chromosome 14 markers to determine the order of contig groups on the chromosome; this should permit us to reduce the number of PCR reactions required for closure of the physical gaps.

To date we have performed 2 cycles of the closure procedure on the chromosome 14 contigs and have nearly completed the third cycle (Table 1). In the first cycle between 12/98 and 2/99, most of our efforts were focused on editing of the contig ends and on performing the sequencing reactions for the missing mates and short sequences identified at physical gaps. The most time consuming and labor-intensive part of the process is editing. Three individuals spent 6 weeks editing the ends of contigs from the initial assembly in order to remove untrimmed vector and low-quality sequences that prevented the merging of overlapping contigs. In subsequent rounds of closure we have re-sequenced an additional 1412 missing mates and short sequences from sequence gaps and have performed 755 primer walks. Between 12/98 and 7/99 we closed 47% of the physical gaps and 65% of the sequence gaps, and one-fourth of the chromosome is now covered by contigs larger than 100 kb. About one-third of the primer walks have yet to be completed and additional editing is underway. Once these steps are completed another assembly will be performed. We expect that > 80% of sequence gaps will have been closed at this point. The remaining gaps are likely to be composed of very AT-rich sequence; closure of these AT-rich gaps will require use of the transposon insertion technique that was used for closure of AT-rich gaps in chromosome 2.

As shown in Table 1, closure of physical gaps has lagged behind closure of the sequence gaps. This is primarily due to the fact that most of our work, apart from the use of database queries to identify missing mates and short sequences at physical gaps, has focused on closure of the sequence gaps. This was done in order to obtain larger contigs that could be placed more accurately on the YAC, SSLP, and optical restriction maps of the chromosome. By locating groups of contigs on the chromosome map PCR reactions using primers from adjacent groups can be used to close physical gaps. About one-third of the physical gaps on chromosome 2 were closed in this way. Once these gaps are closed the remaining gaps can be closed by performing a series of combinatorial PCRs using one primer from a mapped group and another primer from an unmapped group.

Table 1. Progress in gap closure of *P. falciparum* chromosome 14.

	12/98	2/99	3/99	7/99
Sequences	74,292	74,994	75,929	76,406
Contigs	1,750	1,555	1,466	1,418
Largest contig (kb)	99	124	124	164
Total groups	458	293	291	ND[a]
Chr 14 groups	63	37	34	ND
Cum. Length (Mb)	2.99	3.49	3.45	ND
Physical gaps	62	36	33	ND
Sequence gaps	184	180	112	~ 64

[a] ND, not determined.

Recently, however, we prepared primers for all of the physical gaps and have performed PCR reactions with the primers and genomic DNA in order to span these gaps. So far, by performing PCR reactions with primers from the ends of adjacent groups on the chromosome, we have obtained products spanning about 75% of the physical gaps and are in the process of sequencing these products. Many of these PCR products are very AT-rich and have been difficult to sequence. As was done with chromosome 2, many of these PCR products may need to be cloned and subjected to the transposon insertion protocol in order to obtain good sequence data in the AT-rich areas. To obtain PCR products from the remaining physical gaps we have begun a combinatorial PCR procedure in which a primer from one end of a mapped group is tested in series of PCR reactions with primers from the ends of unmapped groups. This process has already generated several new PCR products that are currently being sequenced. We are also investigating use of a multiplex PCR strategy in which pools of four or more primers are used in PCR reactions [10]. This reduces the number of PCR reactions that must be performed during closure and has been very successful in accelerating closure of microbial genomes. The AT-richness of *Plasmodium* DNA makes multiplex PCR more difficult than with other genomes, but we recently obtained PCR products for several physical gaps via multiplexing that are being sequenced.

Perhaps the biggest obstacle faced during the closure process is sequencing through long stretches (up to 50 bp) of As or Ts. We and others have found that the sequence quality deteriorates rapidly as the Taq polymerase passes through these homopolymer stretches, such that accurate sequence data is very difficult to attain in these regions. These regions of lower than average sequence quality have the effect of introducing sequence gaps, which in this case are regions of DNA for which good sequence data cannot be attained. The solution we devised in the chromosome 2 project was to use the transposon insertion method to insert primer binding sites into the AT-rich areas. Frequently, by priming the sequencing reaction within or very close to the homopolymer regions, adequate sequence data could be obtained. However, this is a very labor intensive process and entails performing 50-100 sequence reactions for every gap caused by a homopolymer stretch. To try to improve sequencing of these regions, we are currently testing modifications to our standard sequencing reactions, including changes in extension temperatures, nucleotides mixes, salt concentrations, etc. If these simple modifications improve sequence quality in the AT-rich regions, the gap closure process could be accelerated.

Once all gaps have been closed, the sequence will be evaluated with the program check_coverage to ensure that a) all regions of the assembly are covered by at least two shotgun clones, and b) that every base pair in the sequence has been sequenced in both directions with one chemistry, or in one direction with two chemistries. These criteria ensure that the sequence has been assembled correctly and validate individual base calls. The latter criterion is often satisfied by performing 10% of the sequence reactions with dye-primer chemistry. However, given the frequency of sequence artifacts in AT-rich regions observed with the dye-primer chemistry, this may not be appropriate for *P. falciparum*. As we discovered with chromosome 2, inclusion of sequences containing artifacts in an assembly inhibits contig formation and increases the number of sequence gaps in the assembly and the effort required to close them. Consequently, all chromosome 14 sequencing were done with dye-terminator chemistry. and late in the closure phase the coverage status of the assembly will be assessed. Regions with one-direction coverage will be identified, and additional dye-terminator reactions selected from the database will be performed to convert as many as possible to two-direction coverage. Regions with one-direction coverage that remain and which have unresolved sequence ambiguities will then be re-sequenced with dye-primer chemistry. This process will ensure that the coverage criteria are satisfied and minimize potential assembly problems arising from use of dye-primer chemistry. Finally, the sequence will be edited using the program TIGR_Editor, which displays all gel reads and electropherograms for each base in the sequence. Discrepancies will be noted and additional sequencing reactions will be performed to resolve ambiguities. As a last step to confirm colinearity of the assembled sequence and genomic DNA, restriction maps predicted from the sequence will be compared with the chromosome 14 optical restriction maps described above..

Elucidation of gene structure will be performed with the program GlimmerM, a eukaryotic gene-finding developed at TIGR specifically for the malaria genome project (see section below). Before the annotation of chromosome 14 begins, GlimmerM will be refined to improve accuracy and the training set will be updated with newly-published sequences, so that a more robust gene-finding tool will be available once the sequence is completed. Predicted coding regions will be searched against the sequence and protein databases using our standard methods. Repetitive elements and other features will also be identified and annotated. Since many genes will have no database matches, defining the boundaries of genes will be challenging. Most of the software necessary for annotation was tested during the chromosome 2 project, and will require only a few minor modifications for use on chromosome 14 The annotation performed under this grant will by necessity be preliminary. Our goal is to provide a starting point for further biological characterization. We will facilitate public access to the sequence by release of preliminary and finished sequence on the TIGR web site (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html). This will include full text- and sequence-based searching of chromosomes 2 and 14, as well as links to other sources of *P. falciparum* sequence data such as the Sanger Center and Stanford University. Since the start of the random sequencing phase raw shotgun sequences and contigs from test assemblies have been released on the TIGR web site. Upon completion of the random phase of the project the complete set of > 74,000 shotgun sequences and the contigs from the first full assembly were placed on the web site. These contigs have been updated approximately every 6-8 weeks as gap closure has progressed. In addition, early this year we installed a new BLAST server that returns the BLAST output as well as the FASTA-formatted sequence of the best hit plus 1 kb on either side. This enables

those who are unable to process the very large assembly files to retrieve the sequence of interest without help from the sequencing center.

Sequencing of chromosomes 10 and 11 (Specific Aim 1)

Chromosomes 10 and 11, which together constitute 16% of the genome (1.7 and 2.0 Mb, respectively), are being sequenced primarily with funding provided by the National Institute for Allergy and Infectious Diseases (L.M. Cummings is the Principal Investigator). Funds from this collaborative agreement are being used to accelerate the sequencing, assist in closure and annotation, develop microarrays for these chromosomes, and facilitate rapid utilization of the sequence data by the DoD vaccine and drug development groups. The random phase for chromosomes 11 and 10 was completed in mid- and late- 1999, respectively, and these chromosomes are now in closure. The closure procedure for these chromosomes will be very similar to that used for chromosomes 2 and 14 (see chromosome 14 section above), and will take advantage of any technical improvements that are produced. One major difference in the closure process, however, is that many fewer microsatellite markers are available for these chromosomes[9, 11], making physical gap closure more difficult by reducing the number of contigs that can be accurately ordered on the chromosome. Consequently, Dr. Cummings is collaborating with Dr. X Su of the Laboratory of Parasitic Diseases, NIAID, in production of additional microsatellite markers for these chromosomes. Raw sequence reads and preliminary contigs of chromosomes 10 and 11 have been released on the TIGR web and will be updated periodically as closure proceeds (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html).

Optical mapping of *P. falciparum* chromosomes (added to Specific Aim 1)

Last year we reported that we had collaborated with Dr. David Schwartz in production of optical restriction maps for chromosome 2. We demonstrated that the optical restriction maps were very useful for independent verification of the final chromosome 2 sequence. The successful application of the optical mapping approach to sequence validation of chromosome 2 led the Malaria Genome Sequencing Consortium to recommend that Dr. Schwartz's laboratory be funded by the Burroughs Wellcome Fund to generate two-enzyme restriction maps of the entire *P. falciparum* genome, using high-molecular weight genomic DNA provided by Dr. Dan Carucci, NMRC. A map of the complete genome was recently determined and has been published in Nature Genetics[7]. The whole genome optical map has proven to be very useful in the gap closure process, by assisting in ordering of contigs on the chromosome. In cases where contigs have matches to sequence markers, the optical map provides confirmation of the map position. In cases where contigs do not match any markers, the maps often provide a preliminary localization, which can be confirmed later during gap closure. These optical maps are being used by all of the sequencing centers working on *Plasmodium*, and other investigators working on other parasites are also beginning to use these maps in their genome sequencing projects (e.g. *Trypanosoma cruzi*, N. El Sayed, TIGR, personal communication). Optical maps may prove very useful in the sequencing of other malaria parasites such as *P. yoelii* and *P. vivax*, for which we do not have many sequence tagged sites or microsatellite markers to assist in gap closure. Indeed, preliminary optical map data obtained by Dr. Carucci at NMRC suggests that *P. yoelii*,

contrary to expectations, may have only 13 chromosomes. We are currently working on ways to more efficiently use optical map data in the closure process by writing software that directly maps contigs onto the optical map. Currently this process is done manually.

Development and utilization of a *P. falciparum* gene finding program (added to Specific Aim 2)

Last year we reported the development of new gene finding software, GlimmerM, that was written specifically for annotation of *P. falciparum* chromosome 2. The GlimmerM program [5], when provided with a training set of well-characterized *P. falciparum* genes, constructs a statistical model of coding sequences and donor and acceptor splice sites. New, uncharacterized *P. falciparum* sequence is then analyzed by GlimmerM, and a set of putative gene models is produced. These models are then evaluated by expert annotators in conjunction with other evidence such as database matches, the presence of signal peptides and transmembrane domains, etc., to produce the final gene models reported in the annotation. This year the GlimmerM software has been improved by enlargement of the training set. The size of the training set has been increased by adding 1), gene models from the recently published chromosome 3 sequence [12], 2), other newly published *P. falciparum* sequences from GenBank, and 3), long open reading frames from the chromosome 14 preliminary data. Since GlimmerM works by constructing as statistical model of *P. falciparum* genes, enlargement of the training set should improve the accuracy of the gene predictions. Once chromosomes 10, 11, and 14 have been completed, the first step in the annotation process will be to use the improved GlimmerM software to predict gene models in the chromosome sequence. These gene models will then be searched against the protein and nucleotide sequence databases to identify the genes. GlimmerM has been provided to the other members of the sequencing consortium and is also available from the TIGR web site.

Microarray studies (added to Specific Aim 1)

In last year's report we described our first efforts to add functional genomics studies to this *P. falciparum* genome sequencing project. The aim of these functional genomic studies is to provide a more complete view of *Plasmodium* biology by determining gene expression information for all *Plasmodium* genes that are identifed through the genome sequencing effort. We chose to use glass slide microarrays for this work [13]. Microarrays can be used to examine the expression patterns of thousands of genes simultaneously from two or more RNA samples. These RNA samples may be derived from parasites grown under different growth conditions, or from different life cycle stages, in order to determine the complement of genes that may be differentially expressed under varying conditions. In pilot studies conducted at NMRC to evaulate this technology, PCR products representing virtually all genes from chromosome 2 were prepared and arrayed on glass slides using TIGR's Molecular Dynamics Arrayer robot. Total RNA was prepared from cultured *P. falciparum* (clone 3D7) taken at several time points. The cDNA from each RNA species was differentially labelled with either dUTP-Cy3 or dUTP-Cy5 and hybridized to a DNA microarray at 65° C overnight. After several washes the DNA microarrays were scanned using a ScanArray® 3000 dual color confocal laser system.

Fluorescence intensity measurements of each spot were made using the computer program ImaGene™. Analysis of the data revealed clear examples of differential gene expression during the erythrocytic cycle, which encouraged us to proceed with construction of new microarrays including all genes identified in completed chromosome sequences.

In the past year the microarrays have been expanded to include all genes from chromosome 3 [12], and RNA samples from 6 hourly time points taken throughout the 48 hr erythocytic cycle have been prepared. A series of hybridization experiments have been performed with the chromosome 2 and 3 arrays and the series of erythrocytic stage RNA probes. Analysis of the results is underway and will provide a profile of gene expression throughout the erythrocytic cycle. This information may shed light on the function of the novel genes identified on these chromosomes and and identify potential blood stage antigens for vaccine development. With chromosome 14 now in the late closure phase, we have also begun to prepare primers for amplication of all genes on chromosome 14. When the chromosome 14 PCR products are added to the chromosome 2 and 3 arrays later this year, about 20% of *Plasmodium* genes will be represented on the arrays. As other chromosomes are completed by TIGR/NMRC and other members of the consortium, we will add these genes to the arrays as well.

These experiments produce huge amounts of data. In order to efficiently store and analyze this information, a SyBase relational database was developed at NMRC. SyBase is used at TIGR for storage of all genome sequence and expression data, and so was chosed by NMRC to provide a seamless integration of data generated from the chromosome 2, 10, 11 and 14 projects at TIGR. A Web interface has also been developed to facilitate data entry, tracking, analysis and presentation.

Sequencing of other *Plasmodium* species (Specific Aims 1,2,3)

The primary goal of the Malaria Genome Sequencing Project was to sequence the genome of *P. falciparum*. It appears that within 18 months the random sequencing phase of this project will have been completed, so that virtually all *P. falciparum* genes will have at least partial sequences in the databases. Complete closure of the chromosomes will undoubtably take longer, but malariaologists will nevertheless have access to most *P. falciparum* genes. Even today, with only 2 chromosomes completed, the genome project has had a major effect on malaria research [14-18].

A secondary goal of the project was to sequence the genome of a second species of malaria, and discussions as to which parasite should be chosen had generated lively discussions amongst the malaria community, with some groups favoring seqencing of the human malaria *P. vivax*, and others advocating sequencing one of the rodent malaria parasites that are used as model systems. The sequence of one or more species would be very useful for comparison to P. falciparum, perhaps enabling the identification of genes that may be involved in differences in life cycles and pathogencity, for example. Recent discussions at the semi-annual meeting of the malaria genome consortium may lead to efforts funded by the NIAID, the Burroughs Wellcome Fund, or the Wellcome Trust, to do partial sequencing of several rodent malaria genomes, which will provide useful sequence data to groups working on these different parasites at a reasonable cost.

The TIGR/NMRC team is currently discussing expansion of our sequencing efforts to include *P. vivax*, which is of major concern as a major human pathogen, and the rodent malaria parasite *P. yoelii*, which is used as a model system for vaccine development by NMRC. To date, we have performed limited sequencing of genomic libraries from *P. yoelii*; this data will be compared to the *P. falciparum* genome in order to ascertain how useful the *P. yoelii* data will be for identification of homologs in *P. falciparum*. If successful, it may be possible to rapidly identify P. yoelii homologs of *P. falciparum* vaccine candidates, and facilitate modeling of vaccines in the rodent model. Our primary interest, however, is in sequencing of *P. vivax*, using either a whole genome shotgun approach, or a BAC-based sequence tag connector strategy [19]. Unfortunately, one would strongly prefer to sequence a cloned *P. vivax* parasite and no such clones are available. We are currently trying to determine thefeasibility of cloning this parasite using *in vitro* culture methods or primate models. In the meantime, we have initiated a collaboration with Dr. Thomas Wellems lab at the Laboratory of Parasitic Diseases, NIAID, to sequence a *P. vivax* YAC clone corresponding to the chloroquine resistance locus of *P. falciparum*. Sequencing of this YAC will provide the Wellems lab with important data regarding dchloroquine resistance in *P. vivax*, and also enable us to evaluate the sequencing methods to be used for this genome. We have also arranged a pilot project for production of a *P. vivax* BAC library that will be required for sequencing of the genome.

Modifications to the Specific Aims

Since the random sequencing phase of chromosomes 10, 11, and 14 has been finished and completion of the all 3 chromosomes appears likely by the end of 2001, the TIGR/NMRC team has initiated discussions regarding the best ways to more efficiently use the sequence information to accelerate vaccine development. These discussions are ongoing, but at this point it is clear that information derived from genomics (sequence and annotation), functional genomics (gene expression information), and proteomics (protein expression information) must be applied in a systematic effort to identify candidate vaccine antigens and to evaluate their immunogenicity and protective efficacy. In the next 60 days we expect to have completed an outline of such a program and will take steps to implement it.

Key Research Accomplishments

1) Chromosomes 10, 11, and 14 of *Plasmodium falciparum* have completed the random sequencing phase (over 297,000 sequence reactions) and are now in the gap closure phase.

2) Preliminary sequence data for chromosomes 10, 11, and 14 has been released on the TIGR web site (http://www.tigr.org/tdb/edb/pfdb/pfdb.html).

3) In collaboration with Dr. David Schwartz, an optical restriction map of the *P. falciparum* genome was determined.

4) The GlimmerM gene finding software has been improved by enlargement of the set of sequences used for training of the software

5) Microarrays containing PCR fragments representing virtuall y all chromosome 2 and 3 gene s have been prepared and pilot studies to establish labeling, hybridization, and detection protocols have been completed. A Sybase relational database for storage and analysis of microarray data has been constructed.Experiments to characterize gene expression in erythrocytic parasites are underway.

6) Small-scale projects to sequence portions of the *P. yoelii* and *P. vivax* genomes have been initiated.

Reportable Outcomes

1) Salzberg, S. L., Pertea, M., Delcher, A., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24-31 (1999).
2) Jing, J., Aston, C., Zhongwu, L., Carucci, D. J., Gardner, M. J., Venter, J. C. & Schwartz, D. C. Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Research* **9**, 175-181 (1999).
3) Gardner , M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Smith, H. O., Fraser, C. M., Venter, J. C. & Hoffman, S. L. The malaria genome sequencing project: complete sequence of *P. falciparum* chromosome 2. *Parassitologia* **41**, 69-75 (1999).
4) Gardner, M. J. Invited presentation: The malaria genome project; sequencing of *P. falciparum* chromosome 2., *Universidad de Puerto Rico, Recinto de Ciencias Medicas*, San Juan, Puerto Rico, (1999).
5) Gardner, M. J. Invited presentation: Sequencing of microbial genomes and the implications for vaccine development, *6th Annual IBC Conference of Vaccine Technologies*, Arlington, VA, (1999).
6) Gardner, M. J. Invited presentation: Microbial genome sequencing and vaccine development., *Society for Industrial Microbiology*, Arlington, VA, (1999).
7) Gardner, M. J. Invited presentation: Parasite and fungual genomics at TIGR, *Department of Chemical Engineering, Johns Hopkins University*, Baltimore, MD, (1999).
8) Gardner, M. J. Invited presentation: Microbial genome sequencing and vaccine development, *Society for Industrial Microbiology*, Arlington, VA, (1999).
9) Gardner, M. J. Invited presentation: Malaria research after the genome project, *British Society of Parasitology Malaria Meeting*, Imperial College, London, (1999).
10) Patent application. Chromosome 2 sequence of the human malaria *parasite Plasmodium falciparum* and proteins of said chromosome useful in antimalaria vaccines and diagnostic reagents. Filed by NMRC.

Conclusions

The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic

DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. To date, we have published the first complete sequence of a malarial chromosome (chromosome 2 [4]), completed the random phase sequencing of 3 other large chromosomes totaling 7.2 Mb, and have initiated functional genomics studies using glass slide micorarrays to characterize the expression of chromosome 2, 3, and 14 genes throughout the erythrocytic cycle. We have also collaborated in the construction of a two-enzyme optical restriction map of the entire *P. falciparum* genome [7], and are continuing to further develop the GlimmerM gene finding software developed in year 1. In addition, we have begun small scale sequencing of the rodent malaria *P. yoelii* and are collaborating in the sequencing of part of a *P. vivax* chromosome. Discussions with the Malaria Program, NMRC aimed at development of a program to use genomics and functional genomics to accelerate vaccine research are in progress.

References

1. Organization, W.H., *World malaria situation in 1994: population at risk.* Weekly Epidemiological Record, 1997. **72**: p. 269-276.
2. Butler, D., J. Maurice, and C. O'Brien, *Briefing malaria.* Nature, 1997. **386**: p. 535-540.
3. Bloom, B.R., *A microbial minimalist.* Nature, 1995. **378**: p. 236.
4. Gardner, M.J., *et al.*, *Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum.* Science, 1998. **282**(5391): p. 1126-1132.
5. Salzberg, S.L., *et al.*, *Interpolated Markov models for eukaryotic gene finding.* Genomics, 1999. **59**: p. 24-31.
6. Jing, J., *et al.*, *Optical mapping of Plasmodium falciparum chromosome 2.* Genome Research, 1999. **9**: p. 175-181.
7. Lai, Z., *et al.*, *A shotgun optical map of the entire Plasmodium falciparum genome.* Nature Genetics, 1999. **23**: p. 309-313.
8. Sutton, G.S., *et al.*, *TIGR Assembler: a new tool for assembling large shotgun sequencing projects.* Genome Science and Technology, 1995. **1**: p. 9-19.
9. Su, X.Z. and T.E. Wellems, *Plasmodium falciparum: assignment of microsatellite markers to chromosomes by PFG-PCR.* Exp Parasitol, 1999. **91**(4): p. 367-9.
10. Tettelin, H., *et al.*, *Optimized Multiplex PCR: Efficiently Closing a Whole-Genome Shotgun Sequencing Project.* Genomics, 1999. **62**(3): p. 500-507.
11. Su, X., *et al.*, *A Genetic Map and Recombination Parameters of the Human Malaria Parasite Plasmodium falciparum.* Science, 1999. **286**(5443): p. 1351-1353.
12. Bowman, S., *et al.*, *The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum.* Nature, 1999. **400**(6744): p. 532-8.
13. Brown, P.O. and D. Botstein, *Exploring the new world of the genome with DNA microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 33-7.
14. Jomaa, H., *et al.*, *Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs.* Science, 1999. **285**(5433): p. 1573-1576.
15. Kyes, S.A., *et al.*, *Rifins: A second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 9333-8.
16. Ridley, R.G., *Planting the seeds of new antimalarial drugs.* Science, 1999. **285**: p. 1502-1503.
17. Wellems, T.E., *et al.*, *Genome projects, genetic analysis, and the changing landscape of malaria research.* Curr Opin Microbiol, 1999. **2**(4): p. 415-9.
18. Gardner, M.J., *The genome of the malaria parasite.* Current Opinion in Genetics and Development, 1999. **9**: p. 704-708.
19. Venter, J.C., *et al.*, *Shotgun sequencing of the human genome [see comments].* Science, 1998. **280**(5369): p. 1540-2.

Appendix

1) Salzberg, S.L., *et al.*, *Interpolated Markov models for eukaryotic gene finding.* Genomics, 1999. **59**: p. 24-31.
2) Jing, J., *et al.*, *Optical mapping of Plasmodium falciparum chromosome 2.* Genome Research, 1999. **9**: p. 175-181.
3) Lai, Z., *et al.*, *A shotgun optical map of the entire Plasmodium falciparum genome.* Nature Genetics, 1999. **23**: p. 309-313.
4) Sequencing of malarial parasite genome gets cutting edge boost from optical mapping technique. BioWorld Today, Nov. 4, 1999.
5) Gardner , M.J., *et al.*, *The malaria genome sequencing project: complete sequence of P. falciparum chromosome 2.* Parassitologia, 1999. **41**: p. 69-75.
6) Gardner, M.J., *The genome of the malaria parasite.* Current Opinion in Genetics and Development, 1999. **9**: p. 704-708.

Genomics 59, 24–31 (1999)
Article ID geno.1999.5854, available online at http://www.idealibrary.com on IDEAL®

Interpolated Markov Models for Eukaryotic Gene Finding

Steven L. Salzberg,*,†,[1] Mihaela Pertea,† Arthur L. Delcher,‡,§
Malcolm J. Gardner,* and Hervé Tettelin*

*The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850; †Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218; ‡Department of Computer Science, Loyola College in Maryland, Baltimore, Maryland 21210; and §Celera Genomics, 45 W. Gude Dr., Rockville, Maryland 20850

Received January 19, 1999; accepted April 13, 1999

Computational gene finding research has emphasized the development of gene finders for bacterial and human DNA. This has left genome projects for some small eukaryotes without a system that addresses their needs. This paper reports on a new system, GLIMMERM, that was developed to find genes in the malaria parasite *Plasmodium falciparum*. Because the gene density in *P. falciparum* is relatively high, the system design was based on a successful bacterial gene finder, GLIMMER. The system was augmented with specially trained modules to find splice sites and was trained on all available data from the *P. falciparum* genome. Although a precise evaluation of its accuracy is impossible at this time, laboratory tests (using RT-PCR) on a small selection of predicted genes confirmed all of those predictions. With the rapid progress in sequencing the genome of *P. falciparum*, the availability of this new gene finder will greatly facilitate the annotation process. © 1999 Academic Press

1. INTRODUCTION

The gene finding research community has focused considerable effort on human and bacterial genome sequence analysis. This is not surprising given the attention paid to both areas. The Human Genome Project has produced many millions of nucleotides of sequence, and the importance of rapidly identifying the genes in this sequence cannot be overstated. This task is made difficult by the fact that only 1 to 3% of human genomic sequence is estimated to code for proteins. On the bacterial side, 20 complete bacterial and archaeal genomes have already been published, with dozens more expected in the next 2 years. Gene finders for these prokaryotes have an advantage in that approximately 90% of the DNA of these genomes is coding; thus the task reduces in many cases to choosing between competing reading frames. On the other hand, the demand for accuracy is correspondingly much higher in the prokaryotic world.

[1] To whom correspondence should be addressed. Telephone: (301) 315-2537. Fax: (301) 838-0209. E-mail: salzberg@tigr.org.

In between these two genomic worlds lies a vast array of eukaryotic organisms whose genomes range in size from that of a large prokaryote (on the order of tens of millions of nucleotides) to those that are larger than human (billions of nucleotides). Their gene density tends to be much lower than that of bacteria, but many organisms have a much higher gene density than humans. For example, the genome of the eukaryote *Saccharomyces cerevisiae* has approximately one gene every 5 kb. This corresponds to a gene density of 20%. Recently, chromosome 2 of the malaria parasite *Plasmodium falciparum* was completed (Gardner *et al.*, 1998), and this organism too has a gene density of 20%. The remaining 13 chromosomes from malaria should be completed over the course of the next few years. The much larger (120 million nucleotides) genome of *Arabidopsis thaliana*, which also is expected to have a gene density of approximately 20%, should be completed in the same time frame, and many projects are under way to sequence other small eukaryotes.

Because of their relatively high gene density with respect to human DNA, using a gene finder developed for human sequence (or other organisms with low gene density, including most vertebrates and larger plant genomes) may not be the optimal approach for *P. falciparum* and other small eukaryotes. Prokaryotic gene finders are not well suited to this task because of their inability to handle introns. It is possible to retrain human gene finders using different data (for example, GENSCAN (Burge and Karlin, 1997) has been trained with *Arabidopsis* data), but one still runs the risk that because these systems have been optimized to find genes in DNA that is only 3% coding, they may miss many genes in genomes such as *P. falciparum*.

This paper describes a gene finder developed specifically for small eukaryotes with a gene density of around 20%. This system, GLIMMERM, was built and trained using data from *P. falciparum*, the malaria parasite. It was then used as the principal gene finder for chromosome 2 of *P. falciparum*, which contains 210 genes (209 protein coding genes plus one tRNA) (Gardner *et al.*, 1998). Most of these genes were found by

0888-7543/99 $30.00

GLIMMERM, and as described below, some predictions were confirmed by additional laboratory experiments.

The basis of GLIMMERM is a dynamic programming algorithm that considers all combinations of possible exons for inclusion in a gene model and chooses the best of these combinations. Dynamic programming (DP) has been the basis of many successful eukaryotic gene finders. Hidden Markov model (HMM) systems use a DP algorithm called Viterbi that is a special case of the algorithm here; these HMM methods include VEIL (Henderson *et al.,* 1997); GENSCAN (Burge and Karlin, 1997), which uses semi-Markov HMMs; and Genie (Kulp *et al.,* 1996), which uses generalized HMMs. Very recently, Wirth (1998) described a gene finder for *P. falciparum* based on generalized HMMs, but it is not yet available for comparison. The Morgan system (Salzberg *et al.,* 1996, 1998a) uses a DP algorithm in combination with a decision tree program, and GeneParser (Snyder and Stormo, 1995) uses DP combined with a neural network program. These latter two DP formulations are most similar to the formulation used for GLIMMERM.

2. METHODS AND ALGORITHMS

The phrase "gene model" will be used to denote a particular combination of exons and introns that the system is considering as a possible gene. The decision about what gene model is best is a combination of the strength of the splice sites and the score of the exons produced by an interpolated Markov model (IMM). The methods for producing the IMM and splice site scores are described next, followed by the description of the dynamic programming algorithm that uses these scores.

2.1. Interpolated Markov Models

Markov chains are a family of methods for computing the probability of an event based on a fixed number of previous events. (More formally, a Markov chain is a sequence of random variables X_i, where the probability distribution for each X_i depends only on $X_{i-1}, \ldots, X_{i-k}$ for some constant k.) In the context of DNA sequence analysis, Markov chains predict a base by examining a fixed number of bases just prior to that base in the sequence. The most common type of Markov chain is a fixed-order chain, in which the number of previous bases to examine is a constant. For example, a fifth-order Markov chain will predict a base by looking at the five previous bases. Markov chains, and fifth-order chains in particular, have proven to be effective at gene prediction in bacterial genomes (Borodovsky and McIninch, 1993; Borodovsky *et al.,* 1995).

IMMs are a generalization of fixed-order Markov chains. The main distinction is that rather than deciding in advance how many bases to consider for each prediction, these models will use varying numbers of bases for each prediction. In some contexts they will use 5 bases, while in others they might use 6 or more bases, and in yet other cases they may use 4 or fewer bases. This allows IMMs to be sensitive to how common a particular oligomer is in a given genome. In a given genome, many 5-mers might occur rarely and should not be used for prediction; here the IMM will fall back on a shorter Markov chain. On the other hand, certain 8-mers may occur very frequently, and for those the IMM can use this longer context and make a better prediction. In addition, the IMM can combine the evidence from the eighth-order Markov chain and the fifth-order chain in such cases. Thus it has all the information available to a fifth-order chain plus additional information. It is also worth noting that both IMMs and fifth-order Markov chains should outperform methods based on codon usage statistics. (Cf. Saul and Battistutta (1988), a codon usage method specific to *P. falciparum.* Note that at the time of that work, much less *Plasmodium* data were available, and higher-order statistics might have been inaccurate as a result.)

IMMs form the basis of the GLIMMER system for finding genes in bacteria and archaea (Salzberg *et al.,* 1998b). GLIMMER correctly identifies approximately 98% of the genes in bacteria without any human intervention and with a very limited number of false-positives. It has been used as the gene finder for *Borrelia burgdorferi* (Fraser *et al.,* 1997), *Treponema pallidum* (Fraser *et al.,* 1998), *Chlamydia trachomatis* (Stephens *et al.,* 1998), *Thermotoga maritima* (Nelson *et al.,* submitted for publication), and others. Based on the success of GLIMMER in bacterial sequence annotation, we thought that IMMs should make a good foundation for eukaryotic gene finding. This is particularly true of small eukaryotes like *P. falciparum* in which the gene density is intermediate between that of prokaryotes and higher eukaryotes.

Details of how to construct an IMM for sequence data can be found in the original GLIMMER publication (Salzberg *et al.,* 1998b); GLIMMERM uses the same IMM algorithm as that described there. In brief, GLIMMERM builds IMMs from a set of DNA sequences chosen for training. For coding regions, it builds three separate IMMs, one for each codon position. (This is known as a 3-periodic Markov model (Borodovsky and McIninch, 1993).) These IMMs include zeroth- through eighth-order Markov chains, as well as weights computed for every oligomer of 8 bases or less that appears in the training data. These weights and Markov models are interpolated to produce a score for each base in any potential coding sequence. The logs of these scores are summed to score each coding region.

2.2. Splice Site Identification

The approach used by GLIMMERM to determine the splice sites is similar to that used in the Morgan human gene finding system (Salzberg *et al.,* 1998a). A second-order Markov chain model is used to score a 16-base region around donor sites and a 29-base region around acceptor sites. For both donor and acceptor sites in *P. falciparum,* a wide range of different regions were tested, and these sizes performed best. Two second-order Markov models were built for each type of site. First, a "true" Markov model was created from existing data on known 5′ and 3′ consensus sites. These data were collected by exhaustively combing the literature for every documented exon–intron boundary. A "false" Markov model was built from a large number of randomly chosen false splice sites, i.e., sequences that contained the consensus GT or AG dinucleotide but that were not true splice sites. The score of a site $s_i, s_{i+1}, \ldots, s_j$ was computed by each Markov model according to the formula

$$S(i, j) = \sum_{k=i}^{j} M_{s,k},$$

where

$$M_{s,k} = \ln(f((s_{k-2}, s_{k-1}, s_k), k)/f((s_{k-2}, s_{k-1}), k-1)),$$

and $f(s, k)$ is the frequency of substring s ending at location k. Note that for the leftmost position in the splice site region, M is taken to be the probability given by the zeroth-order Markov model, and for the second position, M is given by the first-order model. The score for a given splice site is computed by taking the difference of the scores obtained from the true site Markov model and the false site model.

After building the models, we scored all the true splice sites and a large selection of randomly chosen false sites. We then set minimum cut-off scores to identify correctly most (or all) true sites and measured how many false-positives we would expect with various thresholds. The splice sites for training the Markov models were taken from the 119 genes (described under Results and Discussion) used to train the IMMs, all of which had laboratory evidence to support them. These genes contained only 81 introns in total, which did not gener-

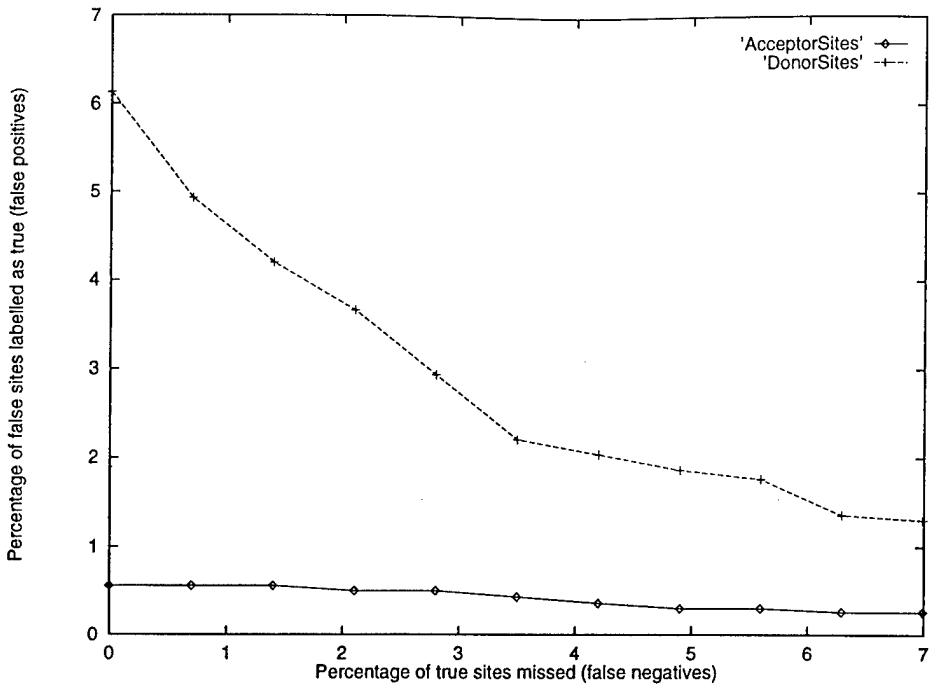

FIG. 1. Trade-off between false-positive rates and false-negative rates for the Markov chain method that recognizes exon–intron splice sites. Data represent the accuracy on sites annotated in chromosome 2 of *P. falciparum*.

ate enough data to produce a very reliable second-order Markov model. Therefore, after an initial training pass using the 81 introns, we used GLIMMERM itself to predict additional introns in chromosome 2, selected the best of these, and added them to the training set. Of course this is a "circular" training protocol, but this represents our attempt to squeeze the best performance we could from limited data. As the sequencing of the remaining chromosomes continues, and as ESTs yield further hard evidence on introns, the available pool of reliable data for training the splice site models should grow dramatically. Alignments with protein sequences from other organisms will provide additional evidence about intron locations. The Markov chain models will consequently improve in accuracy. We intend to continue retraining these models as the genome sequencing progresses.

Figure 1 shows the trade-off in thresholds for the splice site recognition function in *P. falciparum* and shows the trade-off between sensitivity and selectivity for the Markov chain method on the 143 donor and acceptor sites in chromosome 2. Acceptor sites are much easier to recognize: with a false-negative rate of 0% (corresponding to a sensitivity of 100%, meaning that all true sites will be recognized), the false-positive rate—the percentage of AG dinucleotides that will incorrectly be called acceptor sites—is just 0.56%. For donor sites, a 0% false-negative rate corresponds to a rather high 6.1% false-positive rate. Setting the system so that it misses 4 of the 143 (2.8%) donor sites in chromosome 2 would reduce this false-positive rate to 2.9%. The Markov thresholds used here are set so that no true splice sites will be missed.

2.3. Dynamic Programming

GLIMMERM's use of dynamic programming allows it to prune out a large number of possible exon–intron combinations and focus its analysis only on relatively high-scoring combinations (called "parses"). The input to the algorithm is any genomic DNA sequence in FASTA format; small sequences as well as entire chromosomes can be input. The output is a partitioning of the DNA into coding regions interleaved with noncoding regions, on both the main and the complementary strands of the sequence.

As in many other gene finders (Salzberg, 1998), there are a number of assumptions used by GLIMMERM when predicting genes in the DNA sequence. The main assumptions are (1) the coding region of every gene begins with a start codon ATG, (2) a gene has no in-frame stop codons except the very last codon, and (3) each exon is in a consistent reading frame with the previous exon. These constraints significantly enhance the efficiency of computing the optimal gene models, by restricting the search space of the DP algorithm. On the other hand, genuine frameshifts cannot be detected by the system.

The dynamic programming algorithm fills in a structure *Parse*, in which each element *Parse* $[t, n, S]$ denotes the optimal parse of the subsequence that begins at location n and ends at the stop codon at location S. The variable t specifies the type of signal at n, which can be donor, acceptor, start (codon), or stop (codon). More specifically, *Parse* is an ordered list of labeled positions indicating the end-points of a set of exons. For example,

$$Parse[\text{start}, 100, 540] = \langle \text{start}, 100\rangle, \langle \text{donor}, 240\rangle, \langle \text{acceptor}, 380\rangle, \langle \text{stop}, 540\rangle$$

indicates a pair of exons at positions [100 . . . 239] and [380 . . . 539]. A complete gene model is represented as a list *Parse* [start, n, S]. Other elements are partial parses, beginning at a location of type t ($t \neq$ start) and ending at a stop codon S.

The DP algorithm processes the input sequence left to right, looking for stop codons. At each stop codon S, it searches back in the 5′ direction and finds all possible genes ending at that stop. It chooses the highest scoring gene to store in *Parse*. More concisely,

$$Parse[t, n, S] = \langle t, n\rangle, Parse[t_{\text{next}}, i, S],$$

where i is the location that achieves the maximum score

$$\max_{n<i<S}\{Score(\langle t, n\rangle, Parse[t_{\text{next}}, i, S])\},$$

and t_{next} is the type logically following the type t in a parse. For example, if t = acceptor, then t_{next} can be either donor or stop. *Score* (*Parse* [t, n, S]) is the score given by the IMMs to the coding region obtained by concatenating all the exons in the parse delimited by

Parse [t, n, S]. For example, if n is an acceptor site, the algorithm considers all sites i that can follow n and chooses the best one. These would include donor sites, if n is the beginning of an internal exon, and stop codons, if n is the final coding exon. Because the algorithm works backward from each stop codon S, the entry *Parse* [t_{next}, i, S] is computed prior to *Parse* [t, n, S]. The only positions that are considered as possible donor and acceptor sites are those that score above the threshold determined by the Markov chains described previously.

The algorithm incorporates special cases for each of the four types t to prune the search space further. These are as follows:

1. If the interval $(n \ldots i)$ is the coding portion of an exon, its IMM score must exceed a fixed, preset threshold.
2. If two internal exons $(n \ldots i_1)$ and $(n \ldots i_2)$ both result in identical IMM scores, choose the one that maximizes the length of the coding part of the parse. Note that this rule makes GLIMMERM prefer longer gene models.
3. If $(n \ldots i)$ is an intron, then its AT content must be at least 70%. This constraint is based on the observation that all *P. falciparum* introns in the training set had an AT content of above 70%, with only 1% of introns having an AT content under 75%. In contrast, *P. falciparum* exons have an AT content of 70–75%.
4. The length of an intron must be between 50 and 1500 bp; 73 and 1066 bp were the extreme lengths for the introns in the training set.
5. The total length of the coding portions of a gene model represented in *Parse* [start, n, S] must be greater than 200 bp.
6. If n is a stop codon, the algorithm searches backward for all gene models ending at n. Many stop codons can be quickly eliminated because they follow too closely another stop codon in the same reading frame. Thus there is no way to create a gene model ending at these stops—any genes ending at the stop would be too short. The high AT content of *P. falciparum* and the resulting high frequency of stop codons make this step particularly effective.

An attempt was made to use IMMs to score introns as well as exons, but this did not improve the results. Therefore, when t is a donor site and t_{next} is an acceptor, we have

$$Score(\langle \text{donor}, n\rangle, Parse[\text{acceptor}, i, S]) = Score(Parse[\text{acceptor}, i, S]).$$

The algorithm is run separately on both the direct and the complementary strands of the input. GLIMMERM then makes one more pass over the list of putative genes to reject overlapping genes. If genes overlap by less than a fixed amount (30 bp by default), then the overlap is ignored, and both genes are reported in the output. Most overlapping genes are competing gene models that share a stop codon and have different exon locations. Genes that overlap by more than 30 bp are rescored using the IMM, and the gene with the best score is retained. If the scores of two or more overlapping models differ from the maximum score by less than a small preset amount, then GLIMMERM considers the scores equivalent and outputs all the models as possible genes. In these instances, it marks the longest gene as the preferred model.

2.4. Code Availability

The complete GLIMMERM system is available from the authors; it has already been shared with other malaria genome sequencing centers. The code includes routines for retraining the system on data from other organisms. A version of the system trained on *A. thaliana* genes is currently under development. Total processing time to find all genes in malaria chromosome 2 (approximately one million nucleotides) is about 50 min on a Pentium 450 processor running Linux.

2.5. Annotating a Genome

In its current form, GLIMMERM produces multiple gene models for some genes. When no database matches and no other computational evidence were found to support a GLIMMERM prediction, the chromosome 2 annotation reflects the highest scoring model. Although many of these are likely to be correct, it is undoubtedly the case that some are not. Further investigation is required to confirm these predictions (but see below for laboratory evidence confirming a small subset).

The GLIMMERM algorithm was used as one of a suite of tools. Accurate gene identification depends on using every tool available, and the description here should not be taken as implying that GLIMMERM alone can find all genes in *P. falciparum* or any other genome. However, it was a central component in a larger strategy. Other important computational tools used by the malaria chromosome 2 team were as follows: (1) searches of a nonredundant protein sequence database using gapped BLAST and PSI-BLAST (Altschul *et al.*, 1990, 1997); (2) gapped alignments of DNA to protein and EST sequence databases using DDS and DPS (Huang *et al.*, 1997); (3) prediction of putative signal peptides using SignalP (Nielsen *et al.*, 1997); (4) prediction of transmembrane domains with PHThtm (Rost *et al.*, 1995); (5) prediction of nonglobular structures with SEG (Wootton and Federhen, 1996); and (6) a graphical tool to allow annotators to view all the evidence together. In addition, the project used additional aligment tools developed at The Institute for Genomic Research to detect frameshift errors: these tools allow an annotator to detect when a sequence alignment extends beyond the start and stop codons indicated by other tools. In some cases this indicates errors in sequencing, which can be corrected; in other cases it indicates either a genuine frameshift that occurs during translation or a mutation that has changed the length of the translated protein. Any comprehensive annotation effort needs these computational tools and more to produce reasonably accurate gene annotations.

3. RESULTS AND DISCUSSION

GLIMMERM was used as the primary gene finder for chromosome 2 of *P. falciparum*. Chromosome 2 has 209 protein-coding genes spread over approximately one million bases, for a gene density of one gene per 4.5 kb (1/4.5 kb). This contrasts with a density of 1/kb in bacteria, 1/2 kb in yeast, 1/7 kb in *C. elegans*, and 1/50 kb (estimated) in human. Of the 209 protein-coding genes, 43% had at least one intron, and those genes with introns usually had just one or two introns (Gardner *et al.*, 1998). Below we attempt to quantify GLIMMERM's accuracy on these genes.

3.1. Training

To train the IMM, we needed to collect as much coding sequence as possible from *P. falciparum* itself. We exhaustively surveyed the literature to collect every complete sequence that was backed by laboratory evidence. Our survey collected 119 complete coding sequences from 108 GenBank entries representing all 14 chromosomes, of which just 6 genes came from chromosome 2. (This database is available by e-mail upon request from the authors.) Note that by length, chromosome 2 comprises approximately 3% of the genome, so it is unsurprising that just 6/119 genes were from chromosome 2. GenBank contains more than 108 entries from *P. falciparum*, but other entries do not have clear evidence supporting their splice sites. This training set provided the initial data for the splice site models as well.

An important point to emphasize here is that *P.*

TABLE 1

Performance of GLIMMERM on Genes Whose Structure Is Completely Known from Independent Laboratory Evidence

Name	Len	Intr	Comment	Common name
PFB0100c	654	1	Perfect match	Knob-associated His-rich prt
PFB0295w	471	0	Perfect match	Adenylosuccinate lyase (OO)
PFB0300c	272	0	Perfect match	Merozoite surface antigen MSP-2
PFB0305c	272	1	Perfect match	Merozoite surface antigen MSP-5 (EGF domain)
PFB0310c	272	1	Perfect match, highest score from 5 models	Merozoite surface antigen MSP-4 (EGF domain)
PFB0340c	997	3	Perfect match, second highest score from 4 models	SERA antigen/papain-like Protease with active Ser
PFB0405w	3135	0	Perfect match, higher score from 2 models	Transmission blocking Target antigen PfS230

Note. All seven genes had perfect matches to the system's predictions, meaning that the start codon, stop codon, and every splice site were correctly predicted. The column headings give the gene name, its length in amino acids, number of introns (Intr), a comment on GLIMMERM's prediction, and the common name of the protein.

falciparum has an unusually high 82% AT content. As a consequence of this high AT content, stop codons are very frequent (e.g., TAA will occur especially often) in noncoding DNA. This makes it much more likely that long open reading frames (ORFs) represent coding sequence. This fact was used to generate additional training data for GLIMMERM: ORFs greater than 500 bp in the chromosome 2 sequence were assumed to be coding regions and were used in the IMM training. These were added to the list generated by the literature search.

3.2. Accuracy on Known Genes

The 209 genes included in the chromosome 2 annotation were found with GLIMMERM's help. To evaluate the accuracy of the system, it is helpful to consider only those genes from this set for which independent evidence can be found to confirm their existence.

The best way to measure the program's accuracy is to consider its accuracy on those proteins whose exon–intron structure is known precisely from laboratory studies. There are seven genes from chromosome 2 of *P. falciparum* that currently fit into this category; i.e., the sequence from start to stop has been completely characterized. Of these seven, six were included in the training set, and one (PFB0100c) was not.

GLIMMERM's performance on this small set of genes is shown in Table 1. For the two-exon gene PFB0100c, the only independently confirmed gene that was not included in the training set, the system predicted only one model: the correct one. For all seven of the genes, GLIMMERM's output contained a model that matched perfectly. For four of the genes, the correct model was the only one output by the system. For PFB0310c and PFB0405c, GLIMMERM produced five and two competing models, respectively, but in each case the highest scoring one was correct. Only for PFB0340c, a four-exon gene, was GLIMMERM's correct model not the highest scoring one. The system gave a slightly higher score to a model that used a different donor site for the first exon. GLIMMERM's alternate prediction would have a 23-aa insertion in this 997-aa protein.

3.3. Laboratory Tests

An ideal way of measuring the accuracy of GLIMMERM precisely would be to test each of its predictions in the laboratory to see whether they are expressed as predicted. Although a complete test of all predictions would be difficult and time-consuming, one careful set of experiments was conducted as part of the chromosome 2 study.

Because many of the proteins predicted by GLIMMERM had unusual nonglobular domains, the chromosome 2 project team ran a reverse transcriptase (RT-PCR) experiment for 13 of these genes (Gardner *et al.*, 1998) to determine whether or not they were real. These genes are shown in Table 2. The RT-PCR focused its attention on nonglobular domains, not entire proteins, so it could not confirm every detail of the GLIMMERM predictions. In particular, it did not test the exon–intron boundaries for the two genes in this set

TABLE 2

The Set of Genes with Nonglobular Domains for Which RT-PCR Experiments Were Conducted to Confirm Expression

Name	Length	Intr	Common name
PFB0130w	538	0	Prenyl transferase
PFB0145c	1979	0	Hypothetical protein
PFB0180w	560	1	prt with 5′-3′ exonuclease domain
PFB0265c	1516	0	RAD2 endonuclease
PFB0380c	2010	0	Phosphatase (acid phosphatase family)
PFB0435c	1138	7	Predicted amine transporter
PFB0500c	235	0	RAB GTPase
PFB0520w	1233	0	Novel protein kinase
PFB0525w	610	0	Asparaginyl-tRNA synthetase
PFB0685c	885	0	ATP-dependent acyl-CoA synthetase
PFB0720c	899	0	Ori. recognition complex subunit 5 (ATPase)
PFB0755w	1398	0	Hypothetical protein
PFB0880w	426	0	FAD-dependent oxidoreductase

Note. Length is shown in amino acids, and Intr gives the number of introns. In the two genes containing introns, the nonglobular domains are contained within exons.

that contain introns, because the nonglobular domains in those genes do not cross those boundaries. This experiment confirmed that all 13 of the nonglobular domains are expressed; i.e., the predictions for those regions were correct. To our knowledge, this is the first time ever that computational gene predictions provided the impetus for experiments that in turn confirmed the predictions.

Eleven of these 13 genes have sequence homology to known proteins from other organisms. It is worth noting that the nonglobular domains of the *P. falciparum* proteins did *not* occur in the homologs. For example, PFB0180c contains a 176-amino-acid nonglobular insert that is absent from four homologous bacterial exonuclease domains (shown in Fig. 2 of Gardner *et al.*, (1998)). GLIMMERM's prediction for this gene was confirmed by amplifying and then sequencing a region that contained the nonglobular domain. This example points out that the presence of a homologous protein sequence does not always produce an accurate gene prediction.

3.4. Comparison on Genes with Homologs

Of the 209 genes in chromosome 2, 119 have homologous proteins in the public sequence databases. (The training set also contained 119 genes, but the identity of these two numbers is merely coincidence.) The existence of homologs, which come from a wide range of other organisms, provides strong independent evidence that these genes are real. We therefore used these genes to make further measurements of GLIMMERM's accuracy.

Of the 119 genes, 7 were already mentioned: these are the genes from chromosome 2 whose exon–intron structure was known from previously published laboratory studies. Six of those were included in the training set, which leaves 113 genes in chromosome 2 that were *not* included in the training set and for which we have good hints of their exon–intron structure. Because these are homologs, parts of some genes may not align well, making the predicted exon–intron structure less certain.

GLIMMERM finds 98 of these 113 genes (87%) exactly; i.e., the positions of the start codon, the boundaries of each exon and intron, and the stop codon correspond to what is indicated by the alignments to homologous genes. Of these, 22 have competing gene models that score higher, meaning that a human annotator had to examine the output and decide, based on the alignment, to use a model other than the highest-scoring one.

Of the 15 genes that GLIMMERM did not find exactly, 14 were found but had slightly modified coding regions. Seven intronless genes were predicted with incorrect start codons. Three 2-exon genes were broken into two genes each. Four 3-exon genes were predicted with an incorrect first exon but correct second and third exons.

Only one of the genes with homologs, ribosomal protein S30, was missed completely; ribosomal proteins often have a strikingly different composition from other genes and are known to be difficult for content-based gene finders to locate. These will not be missed as long as genomic data are searched against databases of known ribosomal proteins.

In summary, chromosome 2 contains 113 genes that were not included in the set of 119 genes used to train GLIMMERM's IMM. Portions of some of these genes, those with ORFs greater than 500 bp, were extracted automatically and added to the IMM; this portion of the training is fully automatic and requires no human intervention. The splice site training also included some data from chromosome 2, as explained above. A similar procedure can be performed on future chromosomes to extract additional splicing data: first use a sequence alignment program to find homologous genes, extract splice sites from those, and add those splice sites to the Markov chain models. This will allow users of the system to improve the system's performance before making a final run on their chromosomes. Assuming this or a similar protocol is followed, the estimates given here should extrapolate reliably to those chromosomes. Of the 113 genes with homologs, GLIMMERM is able to annotate automatically 76 (66%) if its top-scoring prediction is assumed correct. If a human annotator is available to confirm or reject predictions, then this number grows to 87% (98/113). In most cases the differences between competing models are small, involving one splice site or the start codon. Information from alignments or from other programs—for example, identification of signal peptides—allowed the human annotators to override GLIMMERM's first choice in selected cases.

3.5. Comparison to Chromosome 2 Annotation

Of the 209 genes currently annotated for chromosome 2, GLIMMERM finds 178 exactly. Of these, 40 have competing gene models that score higher; human annotators chose a different model for the final annotation. Of the remaining 31 genes, GLIMMERM finds the stop codons correctly for 14. Different starts appear in the final annotation for several reasons, for example, the existence of a match to a protein sequence that starts at a different start codon. (Note that it is possible that GLIMMERM is still correct in these cases.) The system finds the correct start but the wrong stop codon for 4 genes; this occurs in multiexon genes in which a splice site was missed and one of the exons was incorrectly extended until it hit a stop codon. The 11 remaining partial hits are cases for which GLIMMERM predicts some but not all exons correctly; for example, several multiexon genes are each broken into two separate genes.

Only 2 of the 209 genes are missed completely. One is ribosomal protein S30, which was mentioned above. The second is a predicted integral membrane protein of 192 aa predicted by a preliminary version

of GLIMMERM (before retraining the splice site models). A separate program was used to predict the function of this protein; it did not align to any known sequences.

The improved splice site Markov models resulted in GLIMMERM's generating 41 fewer gene models than before. In addition to the one missed gene just described, it generated 5 new gene models. Of these, one appears to encode a genuine protein, and we are currently investigating this to see if it should be added to the published annotation.

A significant caveat to include with these results is that GLIMMERM often produces multiple competing models that the human annotator must resolve. Most genes with three or more exons result in multiple models. The system indicates which model scores the highest, but as indicated above, 40 of the "correct" gene models had alternative parses that scored higher. These alternative parses share some exons but use different splice sites for others. A human annotator looking at additional evidence, such as alignments to homologous proteins or predictions of signal peptides, was able to overrule the system's top choice in these cases. It is likely that in other cases where no evidence besides GLIMMERM's prediction is available, some of the published annotation may still be in error (all such proteins are annotated as hypotheticals). After each set of multiple gene models was collapsed into one model, the gene list still contains 266 genes. (All of the models can be downloaded on the Web at www.tigr.org/~salzberg/GlimmerMchr2output.html.) These means that, since only 209 genes appeared in the final annotation, the annotators eliminated another 57 gene models entirely from the output. These decisions were somewhat subjective: frequently the putative genes were short or they consisted mostly of low-complexity sequence, and this was not enough to convince the human annotators that the genes were real. In many cases the annotators are probably correct, but it is simply impossible at this point to say with confidence that all of the deleted genes are false-positives. Only further evidence will allow us to decide, but this makes clear the importance of continuing to update and improve genome annotation over time.

ACKNOWLEDGMENTS

S.L.S. is supported by the National Human Genome Research Institute at NIH under Grant K01-HG00022-1. S.L.S., A.L.D., and M.P. are supported in part by the National Science Foundation under Grant IRI-9530462. M.J.G. and H.T. were supported by a supplement to NIAID Grant R01 AI40125-01, which was made possible with funds from NIH's Office for Research on Minority Health and Department of the Army Cooperative Agreement DAMD17-98-2-8005.

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res.* **25**(17), 3389–3402.

Borodovsky, M., and McIninch, J. (1993). Genemark: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17**(2), 123–133.

Borodovsky, M., McIninch, J., Koonin, E., Rudd, K., Medigue, C., and Danchin, A. (1995). Detection of new genes in the bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.* **23,** 3554–3562.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94.

Fraser, C., Casjens, S., Huang, W., Sutton, G., Clayton, R., Lathigra, R., White, O., Ketchum, K., Dodson, R., Hickey, E., Gwinn, M., Dougherty, B., Tomb, J.-F., Fleischmann, R., Richardson, D., Peterson, J., Kerlavage, A., Quackenbush, Salzberg, S., Hanson, M., van R., Vugt, Palmer, N., Adams, M., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M., Horst, K., Roberts, K., Hatch, B., Smith, H., and Venter, J. (1997). Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi. Nature* **390**(6660), 580–586.

Fraser, C., Norris, S., Weinstock, G., White, O., Sutton, G., Clayton, R., Dodson, R., Gwinn, M., Hickey, E., Ketchum, K., Sodergren, E., Hardham, J., McLeod, M., Salzberg, S., Khalak, H., Weidman, J., Howell, J., Chidambaram, M., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M., Horst, K., Roberts, K., Hatch, B., Smith, H., and Venter, J. (1998). Complete genomic sequence of *Treponema pallidum,* the syphilis spirochete. *Science* **281,** 375–388.

Gardner, M., Tettelin, H., Carucci, D., Cummings, L., Aravind, L., Koonin, E., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai, Z., Schwartz, D., Pertea, M., Salzberg, S., Zhou, L., Sutton, G., Clayton, R., White, O., Smith, H., Fraser, C., Adams, M., Venter, J., and Hoffman, S. (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum. Science* **282,** 1126–1132.

Henderson, J., Salzberg, S., and Fasman, K. (1997). Finding genes in human DNA with a hidden Markov model. *J. Computat. Biol.* **4**(2), 127–141.

Huang, X., Adams, M., Zhou, H., and Kerlavage, A. (1997). A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37–45.

Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *In* "ISMB-96: Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology," pp. 134–141, AAAI Press. Menlo Park, CA.

Nelson, K., Clayton, R., Gill, S., Gwinn, M., Dodson, R., Haft, D., Hickey, E., Peterson, J., Nelson, W., Ketchum, K., McDonald, L., Utterback, T., Malek, J., Linher, K., Garrett, M., Stewart, A., Cotton, M., Pratt, M., Phillips, C., Richardson, D., Heidelberg, J., Sutton, G., Fleischmann, R., White, O., Salzberg, S., Smith, H., Venter, J., and Fraser, C. Genome sequence of *Thermotoga maritima:* Evidence for lateral gene transfer between archaea and bacteria. Submitted for publication.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**(1), 1–6.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**(3), 521–533.

Salzberg, S. (1998). Decision trees and Markov chains for gene finding. *In* "Computational Methods in Molecular Biology" (S. Salzberg, D. Searls, and S. Kasif, Eds.), pp. 187–203, Elsevier, Amsterdam.

Salzberg, S., Chen, X., Henderson, J., and Fasman, K. (1996). Finding genes in DNA using decision trees and dynamic programming. *In* "ISMB-96: Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology" pp. 201–210, AAAI Press, Menlo Park, CA.

Salzberg, S., Delcher, A., Fasman, K., and Henderson, J. (1998a). A decision tree system for finding genes in DNA. *J. Computat. Biol.* **5**(4), 667–680.

Salzberg, S., Delcher, A., Kasif, S., and White, O. (1998b). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**(2), 544–548.

Saul, A., and Battistutta, D. (1988). Codon usage in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **27,** 35–42.

Snyder, E. E., and Stormo, G. D. (1995). Identification of coding regions in genomic DNA. *J. Mol. Biol.* **248,** 1–18.

Stephens, R., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R., Zhao, Q., Koonin, E., and Davis, R. (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**(5389), 754–759.

Wirth, A. (1998). "A *Plasmodium falciparum* genefinder," Honours thesis, Department of Mathematics and Statistics, University of Melbourne.

Wootton, J., and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266,** 554–71.

Optical Mapping of *Plasmodium falciparum* Chromosome 2

Junping Jing,[1] Zhongwu Lai,[1] Christopher Aston,[1] Jieyi Lin,[1] Daniel J. Carucci,[2] Malcolm J. Gardner,[3] Bud Mishra,[4] Thomas S. Anantharaman,[4] Hervé Tettelin,[3] Leda M. Cummings,[3] Stephen L. Hoffman,[2] J. Craig Venter,[3] and David C. Schwartz[1,5]

[1]W.M. Keck Laboratory for Biomolecular Imaging, New York University, Department of Chemistry, New York, New York 10003 USA; [2]Malaria Program, Naval Medical Research Institute, Rockville, Maryland 20852 USA; [3]The Institute for Genomic Research, Rockville, Maryland 20850 USA; [4]Courant Institute of Mathematical Sciences, New York University, Department of Computer Science, New York, New York 10012 USA

Detailed restriction maps of microbial genomes are a valuable resource in genome sequencing studies but are toilsome to construct by contig construction of maps derived from cloned DNA. Analysis of genomic DNA enables large stretches of the genome to be mapped and circumvents library construction and associated cloning artifacts. We used pulsed-field gel electrophoresis purified *Plasmodium falciparum* chromosome 2 DNA as the starting material for optical mapping, a system for making ordered restriction maps from ensembles of individual DNA molecules. DNA molecules were bound to derivatized glass surfaces, cleaved with *Nhe*I or *Bam*HI, and imaged by digital fluorescence microscopy. Large pieces of the chromosome containing ordered DNA restriction fragments were mapped. Maps were assembled from 50 molecules producing an average contig depth of 15 molecules and high-resolution restriction maps covering the entire chromosome. Chromosome 2 was found to be 976 kb by optical mapping with *Nhe*I, and 946 kb with *Bam*HI, which compares closely to the published size of 947 kb from large-scale sequencing. The maps were used to further verify assemblies from the plasmid library used for sequencing. Maps generated in silico from the sequence data were compared to the optical mapping data, and good correspondence was found. Such high-resolution restriction maps may become an indispensable resource for large-scale genome sequencing projects.

Optical mapping is a system for the construction of ordered restriction maps from single molecules (Schwartz et al. 1993; Anantharaman et al. 1997). Individual DNA molecules bound to derivatized glass surfaces and cleaved with restriction enzymes are imaged by digital fluorescence microscopy. Resulting cut sites are visualized as gaps between cleaved DNA fragments, which retain their original order (Cai et al. 1995, 1998). Optical mapping has been used to prepare maps of a number of large insert clone types such as bacterial artificial chromosomes (Cai et al. 1998) and most recently genomic DNA (J. Lin, R. Qi, C. Aston, J. Jing, T.S. Anantharam, B. Mishra, D. White, J.C. Venter, and D.C. Schwartz, in prep). A shotgun mapping strategy was developed in parallel for several microorganisms using large fragments of randomly sheared DNA that were mapped with high cutting efficiencies. The numerous overlapping restriction site landmarks and a measurable cutting efficiency combined together to enable accurate contig assembly without the use of cloned DNA (Anathraman et al. 1998). Because library construction was obviated, it was possible to map large *Plasmodium falciparum* (*P. falciparum*) DNA fragments, which are AT-rich and notoriously difficult to clone because of deletion and rearrangement in *Escherchia coli* (Gardner et al. 1998). Because cloning artifacts were precluded, this enabled accurate maps to be generated. Furthermore, small amounts of starting material were used, facilitating the mapping of this and potentially other parasites that are problematic to culture or clone.

Sequencing of chromosome 2 of *P. falciparum* was completed recently by Gardner and colleagues (Gardner et al. 1998), as part of an international consortium sequencing the whole *P. falciparum* genome (Foster and Thompson 1995; Dame et al. 1996). Existing physical maps of *P. falciparum* chromosomes [chromosome 3; (Thompson and Cowman 1997) and chromosome 4 (Sinnis and Wellems 1988; Watanabe and Inselberg, 1994)], prepared by restriction digestion, gel fingerprinting, and hybridization of probes are of moderate resolution and not ideally suited for systematic sequence verification. To assess the feasibility of optically mapping a whole eukaryotic chromosome, we constructed high-resolution, ordered restriction maps of *P. falciparum* chromosome 2 using genomic DNA and later compared these maps with those generated in

[5]Corresponding author.
E-MAIL schwad01@mcrcr.med.nyu.edu; FAX (212) 995-8487.

silico from the sequence data. Such restriction maps reveal the architecture of large spans of the genome and have value in shotgun sequencing efforts because they provide ideal scaffolds for sequence assembly, finishing, and verification. Gaps that form between contigs can be characterized in terms of location and breadth, thereby facilitating closure techniques.

RESULTS

P. falciparum Chromosome 2 DNA Sample

A chromosome 2 gel slice was used as starting material. Despite the AT-rich nature of the *P. falciparum* genome (80–85%), melting of low-gelling-temperature agarose inserts did not affect the integrity of the DNA and the chromosomal DNA was competent for optical mapping. Previously, we mounted DNA molecules by sandwiching the sample between an optical mapping surface and a microscope slide, followed by peeling the surface from the slide. DNA molecules were stretched and fixed onto the surface. This method works very well with clone types such as bacteriophage, cosmid, and BAC (Cai et al. 1995, 1998); however, larger genomic DNA molecules tend to form crossed molecules. We improved this approach by adding the sample to the edge formed by the placement of a surface onto a slide. The liquid DNA sample spreads into the space between the surface and the slide by capillary action. Consequently, DNA breakage was minimized, molecules tended to elongate in the same direction, and crossed molecules were also minimized (see Fig. 1).

*Nhe*I and *Bam*HI Maps for *P. falciparum* Chromosome 2

The genomic DNA was mapped with either *Nhe*I (Fig. 1A) or *Bam*HI (Fig. 1B). Fragment sizes were calculated by comparison with comounted λ bacteriophage DNA (48.5 kb). *P. falciparum* DNA has an AT content of 80–85% and λ bacteriophage DNA has an AT content of 50%. The YOYO-1 fluorochrome used for DNA staining intercalates preferentially between GC pairs with increased emission quantum yield (Netzel et al. 1995). A correction factor was therefore applied to each fragment size to correct for this massively different fluorochrome incorporation. λ bacteriophage DNA was used also to determine areas on the surface where cutting efficiency was highest. Cutting efficiencies were > 80%. Maps were obtained from individual molecules of ~350 kb. Consensus maps were assembled from 50 molecules generating an average contig depth of 15 molecules. Chromosome 2 was found to be 976 kb by optical mapping with *Nhe*I, and 946 kb by optical mapping with *Bam*HI (average size 961 kb). There were 40 fragments in the *Nhe*I map, ranging from 1.5–115 kb, with average fragment size 24 kb (Fig. 2). There were 30 fragments in the *Bam*HI map ranging from 0.5–80 kb, with average fragment size 32 kb (Fig. 2). Each fragment size in the consensus map was averaged from 10 to 15 fragments. Although *P. falciparum* chromosome 2 migrates as a distinct band by PFGE, we found the gel slice to contain only 60% chromosome 2-specific DNA. The remaining optical mapping data was rejected.

Integration of Optical Maps and Sequence Data

The chromosome 2 sequence assembled by Gardner and colleagues shows chromosome 2 to be 947 kb (Gardner et al. 1998) versus 976 kb by optical mapping with *Nhe*I and 946 kb with *Bam*HI. The optical restriction maps were compared to restriction maps predicted from the sequence, and there was very good correspondence between the two, indicating that there were no major rearrangements or errors in the assembled sequence (Table 1). The optical map included all fragments above 500 bp predicted from sequence. The overall agreement between these maps and the sequence was therefore excellent, with the average fragment size difference below 600 bp (relative error 4.3%) for the *Nhe*I map. The average fragment size difference for the *Bam*HI map was 1.2 kb (relative error 5.8%). However, there were several notable differences. Large differences in size for the fragments at each end of the chromosome were noted (Tables 1 and 2). This is because the sequence for these subtelomeric regions is still under construction. PCR products spanning subtelomeric gaps are being sequenced currently. The optical map sizes were larger than those predicted from sequence for certain other fragments (Tables 1 and 2). These differences were due to large fluorescence inten-

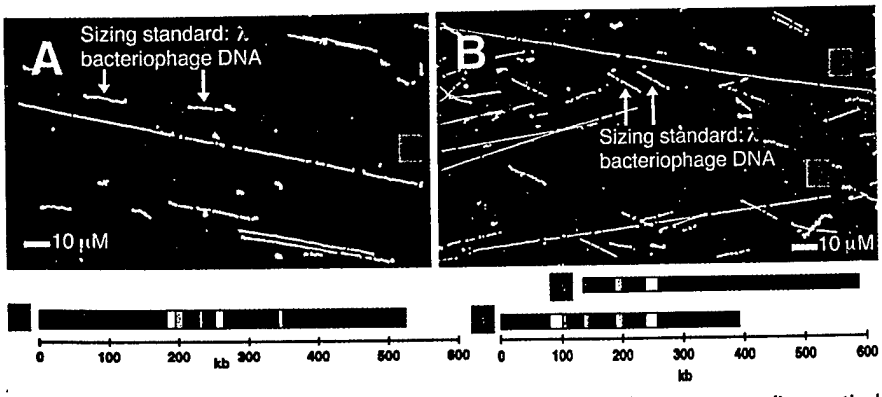

Figure 1 Typical *P. falciparum* chromosome 2 molecules and their corresponding optical maps. (*A*) digested with *Nhe*I (*B*) digested with *Bam*HI. Maps derived from the two *Bam*HI-digested molecules in (*B*) can be aligned.

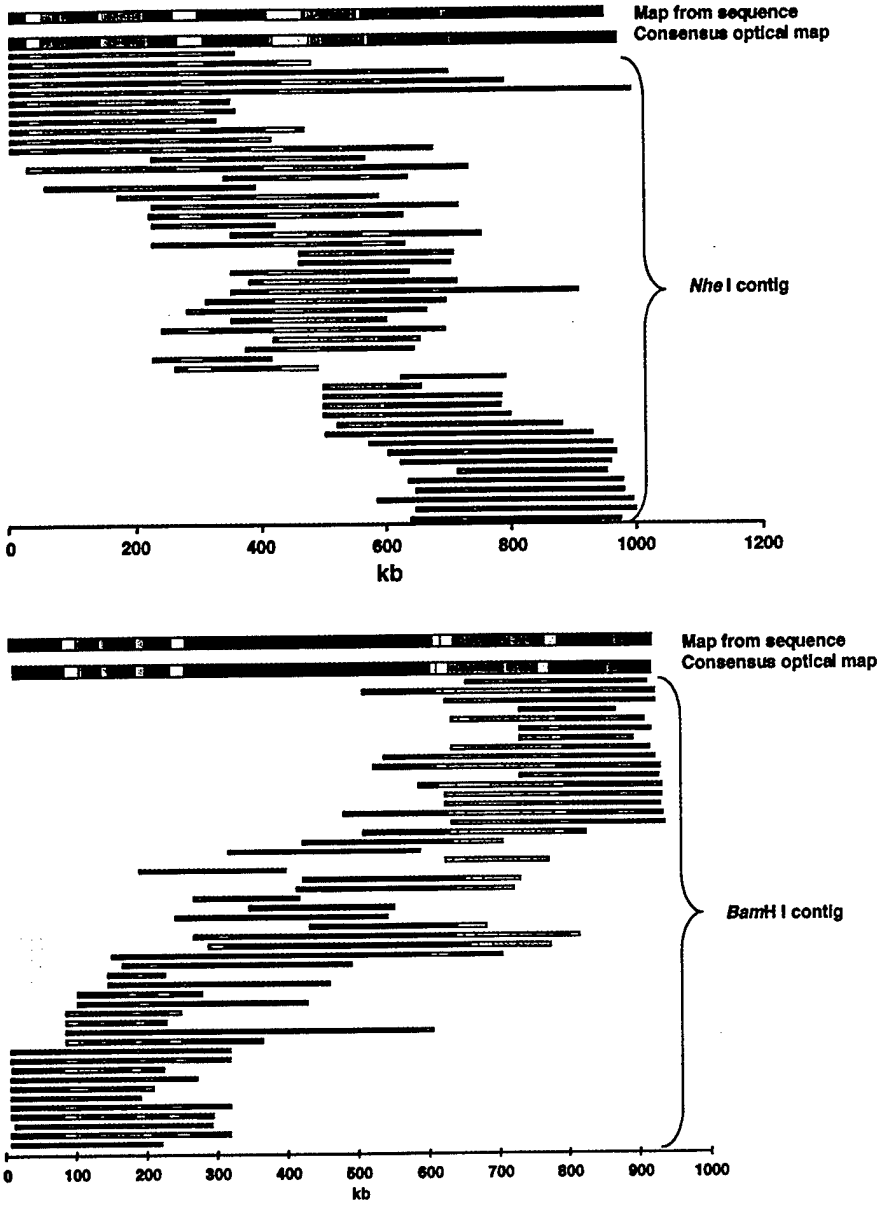

Figure 2 High-resolution optical mapping of *P. falciparum* chromosome 2 using *Nhe*I (*A*) and *Bam*HI (*B*). The underlying contig used to generate the consensus map is shown. The map predicted from sequence information is shown for comparison.

sity measurements falsely caused by crossed molecules. Currently, we combine length measurements with fluorescence intensity measurements to improve on our sizing of these fragments. Chromosome 2 maps using these new measurements show no exceptional errors (not shown; Jing et al., in prep). The map was used to facilitate sequence verification. Optical maps can also be used at the earlier sequence-assembly stage to form a scaffold for assembly of contigs formed from sequencing. Linking of single-enzyme maps produces a much higher resolution multi-enzyme map that is rich in information. Smaller contigs can be placed confidently on a multi-enzyme map. Nowadays, mapping is rarely done in the absence of sequencing. Figure 3 shows a comparison of a multi-enzyme map generated by optical mapping with that predicted from sequence. The maps are in complete agreement across the whole length of the chromosome. Given even small amounts of sequence (~100 kb), maps can be linked and verified readily.

Map Confirmation by Southern Blotting

To confirm the optical maps independently of sequence data, pulsed-field gels of total *P. falciparum* DNA digested with *Nhe*I or *Bam*HI were run and blotted. Plasmid clones used as sequencing templates provided the probes to analyze the Southern blots. Restriction fragment sizes of the blots closely compared in size to the fragments determined by optical mapping and those predicted from the preliminary sequence. Probe PF2CM93 hybridized to a 7.5 kb band generated by *Nhe*I digestion and PFGE. The fragment size predicted from sequence information was 7.6 kb. The corresponding fragment size from the optical map was also 7.6 kb (Table 1). The same probe hybridized to a 41-kb band generated by *Bam*HI digestion and PFGE. The fragment size predicted from sequence information was 41.3 kb. The corresponding fragment size from the optical map was 40.8 kb (Table 2). Probe PF2NA66 also generated data with fragment sizes that were very similar (Tables 1 and 2). By using the same probe on DNA digested with the two different enzymes, the optical maps were oriented and linked with one another.

DISCUSSION

We have generated a high resolution *Nhe*I and the *Bam*HI optical restriction map of *P. falciparum* chromosome 2, which was used in sequence verification. Despite the fact that chromosome 2 is resolved easily by PFGE, we found the chromosome 2 gel slices to contain only 60% chromosome 2-specific DNA. The balance

Table 1. Comparison of *Nhe*I Optical Map with Restriction Map Predicted from Sequence

Optical map (kb)	Map predicted from sequence (kb)	Difference (kb)	Relative difference (%)	Hybridizing probe
71.8	66.597	5.24		
114.5	115.147	0.63	0.6	
10.3	10.226	0.02	0.2	
3.4	3.359	0.07	2.1	
7.9	7.856	0.05	0.6	
24.7	23.684	1.03	4.4	
6.8	4.933	1.88	38.0	
16.5	14.553	1.97	13.6	
3.2	2.875	0.30	10.3	
	0.177			
11.5	11.425	0.10	0.9	
4.1	3.768	0.30	7.9	
63.8	63.252	0.50	0.8	
10.0	10.018	0.01	0.1	
6.7	6.431	0.27	4.2	
8.9	9.248	0.31	3.3	
28.7	27.327	1.34	4.9	
4.3	4.357	0.07	1.6	
7.6	7.581	0.01	0.01	PF2CM93
11.0	10.588	0.44	4.2	
60.5	60.324	0.21	0.4	
12.3	11.935	0.40	3.3	
4.1	3.964	0.12	3.0	
58.2	57.925	0.25	0.4	
5.5	5.381	0.07	1.3	
	0.363			
1.6	1.546	0.02	1.5	
23.4	22.405	0.96	4.3	
35.1	34.171	0.91	2.6	
18.1	17.156	0.93	5.4	
3.1	2.947	0.16	5.4	
24.9	25.138	0.28	1.1	
40.8	40.107	0.73	1.8	
20.8	20.176	0.59	2.9	
25.1	24.476	0.62	2.5	
77.3	75.172	2.15	2.9	PF2NA66
16.6	16.637	0.07	0.4	
48.0	45.683	2.30	5.0	
9.4	8.546	0.88	10.3	
20.1	18.986	1.15	6.0	
23.9	23.192	0.75	3.2	
32.1	14.897	5.65		
976.5	934.513	0.60	4.3	

was contaminated with DNA molecules from other chromosomes. Consequently, a portion of the optical mapping data was rejected. Should we have mapped other chromosomes using the same strategy we could not predict the acquisition of concise data from chromosomes, which are less resolvable by PFGE, such as chromosomes 5–9.

To check the fidelity of the optical maps independently, Southern blotting of chromosome 2 DNA was performed. Sequenced small-insert clones were used as probes, enabling the optical maps to be cross-checked against the sequence. In all, the optical maps were verified against sequence data and Southern blot analysis, and were found to be very accurate. A more directed operation would be to use sequence-templates as probes for hybridizations to generate a series of anchors for sequence assembly. Such templates would be placed precisely onto the optical map, in terms of physical distance (kb) and would be critical for finishing genomic regions of high complexity; namely, tandem or inverted repeats of high homology and short sequence length. This approach would also readily assemble data acquired using different techniques and would allow the placement of very short sequence contigs onto a map. For example, STS markers or ESTs could be assigned to restriction fragments on a whole genome optical map.

Optical maps of entire chromosomes should also find utility at the sequence-assembly stage in which numerous large contigs are formed, but have unknown order along a chromosome. Traditional approaches to establish contig order rely, in part, on combinatorial PCR, or sequence alignment with physical landmarks,

Table 2. Comparison of *Bam*HI Optical Map with Restriction Map Predicted from Sequence

Optical map (kb)	Map predicted from sequence (kb)	Difference (kb)	Relative difference (%)	Hybridizing probe
77.1	76.648	0.42		
19.9	20.955	1.07	5.09	
7.5	6.81	0.65	9.52	
26.1	27.054	0.95	3.52	
9.9	9.831	0.11	1.15	
41.0	43.295	2.28	5.26	
12.4	13.647	1.22	8.92	
3.7	3.754	0.02	0.67	
34.8	35.985	1.18	3.28	
21.1	20.22	0.91	4.51	
63.6	61.785	1.80	2.92	
55.9	55.217	0.73	1.32	
41.3	40.788	0.50	1.22	PF2CM93
67.3	70.318	3.05	4.33	
46.7	46.943	0.23	0.49	
81.2	87.327	6.14	7.03	
2.0	1.786	0.20	11.35	
8.9	11.633	2.68	23.07	
18.6	17.953	0.69	3.85	
80.8	83.96	3.16	3.77	
19.9	20.665	0.78	3.76	
31.1	30.351	0.72	2.39	
17.4	17.959	0.56	3.10	
28.6	30.812	2.22	7.21	PF2NA66
52.2	49.95	2.26	4.52	
2.0	1.813	0.18	9.70	
24.9	24.79	0.07	0.28	
6.0	5.315	0.65	12.28	
0.5	0.621	0.12	19.48	
34.8	16.346	6.93		
937.2	934.531	1.25	5.86	

which are usually well defined in terms of order but not physical distance. This is where optical maps can streamline the final assembly process by reducing the required number of PCR reactions, by providing an easily interpretable physical scaffold with which sequence contigs can be aligned. The alignment process is to simply generate restriction maps in silico from the sequence data and compare this with the optical maps. When multiple enzymes are used independently and resulting maps are aligned properly, the composite map decreases the size of the sequence contig necessary for confident alignment to the final scaffold.

The information content of a multiple restriction enzyme map is greater than the sum of its parts (Lander and Waterman 1988). We used the sequence data to align the *Nhe*I and *Bam*HI restriction maps with respect to each other, creating a composite map. We expected to find a number of restriction site reversals in this composite. That is, given our sizing errors, closely spaced fragments in the composite map may not be represented in the correct order, and would possibly shift relative position. To our initial astonishment, we found only one instance of reversal. Given this result, we decided to evaluate its statistical significance.

One way to evaluate the quality of a composite enzyme map is to examine how well it preserves the order of the restriction sites. For instance, if we create two maps, one with a restriction enzyme A and the other with the restriction enzyme B, and combine the two maps in correct order, it is still possible that the sizing error in the individual fragments may create a situation, in which a restriction site of type B appears before A, whereas the correct order (in the sequence) is A followed by B— restriction sites shift. Assume that both enzymes cut at the same rate E, and the genome (or chromosome) length is L. Then the total number of fragments of each type is $N = LE$. If the sizing error in a fragment is σ (for instance 1 kb), then the maximum sizing error occurs in the middle of the map and is bounded by $(\sqrt{N/2})\sigma$ (a rather conservative estimate). Thus, a fragment of length l, and cuts of type A in one end and of type B in the other end, may appear in the computed map as a fragment whose length is a random variable with mean l and standard deviation $\sigma' = (\sqrt{N})\sigma$. Thus the probability that this fragment will appear in the reversed order is bounded by $\Phi(l/\sigma')$, where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du$$

Furthermore, the length of the fragment with cuts A and B is distributed as $2Ee^{-2El}$. Thus, a random fragment of this kind has a length longer than σ' with probability $e^{-2E\sigma'}$ and a simple estimate shows that the probability of reversal is bounded by

$$(1 - e^{-2E\sigma'})\Phi(0) + e^{-2E\sigma'}\Phi(1)$$

Consider the following values of the parameters $L = 980$ kb, $E = 1/30{,}000$, $\sigma = 1$ kb. For these values, $\sigma' = 5.7$ kb and the average fragment length (with two enzymes) is 15 kb. The above estimate indicates that the probability of reversal is bounded by 0.27. A somewhat better estimate can lower this value to 0.17. As the expected number of fragments with cuts A following B (or B following A) is ~30, one would expect to see fewer than five reversals.

However, the composite map created by optical mapping has only one reversal. The probability of this situation (with fewer than 1 reversal) occurring is ~1 in 40. More exactly, this probability is $(1 - p)^{30} + 30\ p\ (1 - p)^{29} = 0.023$. This difference may signal the requirement for more sophisticated analysis, or indicates the presence of a potentially useful physical effect. A closer examination of the data reveals that the error in the fragment sizes in the composite map has a normal distribution with mean, 0.02 kb and standard deviation, 2.01 kb. Surprisingly, the error in the cut locations has a mean, −1.78 kb and a standard deviation, 1.82 kb, indicative of the presence of systematic (e.g., sequence-specific) error and much smaller unsystematic error. A recalculation of the expected number of reversals with the observed values ($\sigma' = 1.82$ kb) results in slightly more than two reversals, making the observed number of reversals of only one much more likely (~1 in 7 as opposed to 1 in 40). Note that as our estimate of σ' is for the worst-case situation, we believe a more realistic analysis may close the gap. On the other hand, this may be caused by another biochemical effect that we

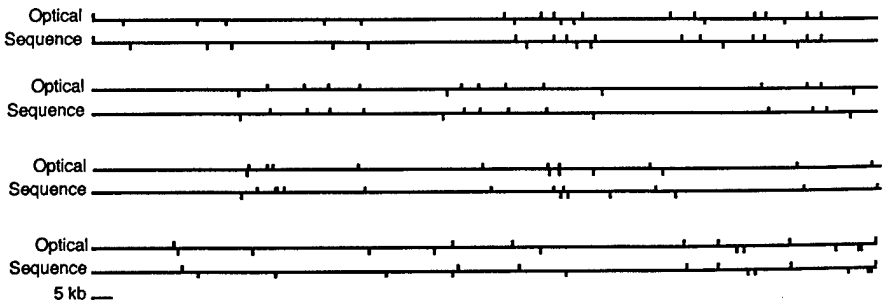

Figure 3 The use of sequence information to link single enzyme maps. The top map was generated by normalizing the single enzyme maps to be the same size (961 kb). The resulting multienzyme map was aligned with the map predicted from sequence. The median relative error is 7%. The average absolute error is 1.4 kb. Upper tick marks are *Nhe*I sites; lower tick marks are *Bam*HI sites.

do not account for in our analysis. More experiments and analyses are required to resolve this situation.

Current optical mapping studies of *P. falciparum* use whole genomic DNA as starting material. The chromosomes are resolved at the level of data rather than as physical entities. The data segregates into 14 deep contigs corresponding to the various chromosomes. Chromosome 2 can be resolved based on size and the near complete correspondence with the data shown in this paper (one 600-bp *Bam*HI fragment is missing on the whole genome map). The success of this project has prompted the Malaria Genome Consortium to recommend support of whole genome mapping to assist in closure of chromosomes, as well as for verification of the final assembly.

In summary, we describe the construction of an ordered restriction map of *P. falciparum* chromosome 2 using optical mapping of genomic DNA. A combined approach using shotgun sequencing and optical mapping will facilitate sequence assembly and finishing of large and complex genomes.

METHODS

Parasite Preparation

P. falciparum (clone 3D7) was cultivated using standard techniques (Trager and Jensen 1976). To minimize possible alterations of the genome that can occur in continuous culture (Corcoran et al. 1986), parasite aliquots were kept frozen in liquid N_2 until needed and then cultivated only as long as necessary. Parasites were cultivated to late trophozoite/early schizont stages and enriched on a Plasmagel gradient. The parasitized red blood cells were washed once with several volumes of 10 mM Tris (pH 8), 0.85% NaCl and the parasites were freed from the erythrocytes by incubation in ice-cold 0.5% acetic acid in dH_2O for 5 min, followed by several washes in cold buffer. The parasites were resuspended to a concentration of 2×10^9/ml in buffer and maintained in a 50°C waterbath. An equal volume of 1% InCert agarose (FMC, Rockland, ME) in buffer, prewarmed to 50°C, was mixed with the prewarmed parasites and the mixture was added to a $1 \times 1 \times 10$-cm gel mold, plugged at one end with solidified agarose, and was allowed to cool to 4°C. The agarose-embedded parasites were pushed out of the mold and incubated with 50 ml of proteinase K solution (2 mg/ml proteinase K in 1% Sarkosyl, 0.5 M EDTA) at 50°C for 48 hr with one change of proteinase K solution and were stored in 50 mM EDTA at 4°C (Schwartz and Cantor 1984).

Chromosome 2 Isolation by PFGE

Uniform parasite slices were taken with a glass coverslip using two offset microscope slides as guides. One half to one quarter of a single slice was sufficient per lane. Parasite slices were arranged end to end on the flat side of the gel comb. The parasites were fixed to the comb by a small bead of molten (60°C) agarose. The comb was then placed into the gel mold and molten agarose [1.2% SeaPlaque (FMC) in 0.5× TBE] poured around the parasite-containing slices. Once cooled, the comb was removed and the space filled with molten agarose. A CHEF DRIII apparatus (Bio-Rad, Hercules, CA) was used for all PFGE (Schwartz and Cantor 1984) chromosome separations. Gels were run with 180–250 sec of ramped pulse time at 3.7 V/cm and 120° field angle, for 90 hr at 14°C with recirculating buffer at ~1 l/min, using *Saccharomyces cerevisiae* and/or *Hansenula wingei* PFGE size markers (Bio-Rad). To minimize UV damage to the DNA, gel slices were removed from the ends of the gel, stained with ethidium bromide (5 µg/ml), and visualized by long wave (320 nm) UV light. Notches corresponding to the individual chromosomes were made in the agarose gel and used as guides to cut the chromosome from the gel. The chromosome-containing gel slices were stored in 50 mM EDTA at 4°C until needed. The gel was stained with ethidium bromide to verify the chromosome excision. The genome of *P. falciparum* is 26–30 Mb in size, consisting of 14 chromosomes ranging in size from 0.6–3.5 Mb (Foote and Kemp 1989). PFGE resolves most of the *P. falciparum* chromosomes, except 5–9 which are similar sizes and comigrate. The gel band containing *Plasmodium falciparum* chromosome 2 was resolved easily, cut from the gel, melted at 72°C for 7 min and incubated with agarose at 40°C for 2 hr. The melted agarose band was diluted in TE to a final DNA concentration suitable for optical mapping (~20 pg/µl.

Mounting and Digestion of DNA on Optical Mapping Surface

Optical mapping surfaces were prepared as described previously (Aston et al. 1999). Briefly, glass coverslips (18×18 mm^2; FISHER Finest, Pittsburgh, PA) were cleaned by boiling in concentrated nitric, then hydrochloric acid. Surfaces were derivatized with 3-aminopropyldiethoxymethyl silane (APDEMS; Aldrich Chemical, Milwaukee, WI). One surface was placed onto a microscope slide. A DNA sample (10 µl) was added to the edge between the surface and the slide and spread into the space between the surface and the slide. The surface was then peeled off from the slide. Digestion was performed by adding 100 µl of digestion solution [50 mM NaCl, 10 mM Tris-HCl (pH 7.9), 10 mM $MgCl_2$, 0.02% Triton X-100, 20 units of restriction endonuclease; New England Biolabs, Beverly, MA] onto the surface and incubating at 37°C from 15 to 30 min. The buffer was aspirated and the surface washed with water before staining of DNA with YOYO-1 homodimer (Molecular Probes, Eugene, OR), prior to fluorescence microscopy. Comounted λ bacteriophage DNA (New England Biolabs) was used as a sizing standard and also to estimate cutting efficiencies.

Image Acquisition, Processing, and Map Construction

DNA molecules were imaged by digital fluorescence microscopy. The optical mapping surface was scanned by the operator for individual digested DNA molecules of adequate length and quality to be collected for image processing and map making. Images were collected with a cooled charge coupled device (CCD) camera (Princeton Instruments, Trenton, NJ) using Optical Map Maker (OMM) software, as described previously (Jing et al. 1998). Images of DNA fragments were processed using a modified version of NIH Image (Huff 1996) which integrates fluorescence intensity for each fragment. These values were used to assemble an ordered restriction map for each molecule. Fluorescence intensity of λ bacteriophage DNA standards was used to measure the size of the *P. falciparum* restriction fragments on a per image basis. Cutting efficiences (on a per image basis) were determined from scoring

cut sites on sizing standard molecules contained in the same field as the genomic DNA molecules. Standard molecules were cut once by *Nhe*I and five times by *Bam*HI. The map for the entire chromosome 2 was manually assembled into contigs by aligning overlapping regions of congruent cut sites. If there were no overlapping regions, the molecules were considered to be from a contaminating *P. falciparum* chromosome and were discarded. Consensus maps for chromosome 2 were assembled by averaging the fragment sizes from the individual maps derived from maps underlying the contigs.

Southern Blotting of *P. falciparum* Genomic DNA

P. falciparum genomic DNA (10 µg) was digested with *Nhe*I or *Bam*HI, resolved by PFGE (POE apparatus, 1% gel in 0.5× TBE, pulse time, 1 sec, 2 sec; switch time, 12 sec, 150 V, for 24 hr) (Schwartz and Koval 1989), blotted, and hybridized with probes derived from small insert clones used for sequencing (PF2CM93 and PF2NA66). Probes were labeled by random priming.

ACKNOWLEDGMENTS

This work was supported by the Burroughs Wellcome Fund and the Naval Medical Research and Development Command work unit STEP C611102A0101BCX. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the U.S. Navy or naval service at large.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ananthraman, T.S., B. Mishra, and D.C. Schwartz,. 1997. Genomics via optical mapping II: Restriction maps. *J. Comput. Bio.* **4:** 91–118.

Anantharaman, T.S., B. Mishra, and D.C. Schwartz. 1998. Genomics via optical mapping III: Contiging genomic DNA and variations. *Courant Technical Report #760*, Courant Institute, New York.

Aston, C., C. Hiort, and D.C. Schwartz. 1999. Optical mapping: An approach for fine mapping. *Methods Enzymol.* **303:** (in press).

Cai, W., H. Aburatani, D. Housman, Y. Wang, and D.C. Schwartz. 1995. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc. Natl. Acad. Sci.* **92:** 5164–5168.

Cai, W., J. Jing, B. Irvin, L. Ohler, E. Rose, U. Kim, M. Simon, and D.C. Schwartz. 1998. High resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proc. Natl. Acad. Sci.* **95:** 3390–3395.

Corcoran, L.M., K.P. Forsyth, A.E. Bianco, G.V. Brown, and D.J. Kemp. 1986. Chromosome size polymorphism in plasmodium falciparum can involve deletions and are frequent in nature parasite populations. *Cell* **44:** 87–95.

Dame, J.B., D.E. Arnot, P.F. Bourke, D. Chakrabarti, Z. Christodoulou, R.L. Coppel, F. Cowman, A.G. Craig, K. Fischer, J. Foster et al. 1996. Current status of the *Plasmodium falciparum* genome project. *Mol. Biochem. Parasitol.* **79:** 1–12.

Foote, S.J. and D.J. Kemp. 1989. Chromosomes of malaria parasites. *Trends Genet.* **5:** 337–342.

Foster, J. and J. Thompson. 1995. The *Plasmodium falciparum* genome project: A resource for researchers. The Wellcome Trust Malaria Genome Collaboration. *Parasitol. Today* **11:** 1–4.

Gardner, M.J., H. Tettelin, D.J. Carucci, L.M. Cummings, L. Aravind, E.V. Koonin, S. Shallom, T. Mason, K. Yu, C. Fujii et al. 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282:** 1126–1132.

Huff, E. 1996. Ph.D. thesis. Department of Chemisty, New York University, New York, NY.

Jing, J., J. Reed, J. Huang, X. Hu, V. Clarke, J. Edington, D. Housman, T. Anantharaman, E. Huff, B. Mishra et al. 1998. Automated high resolution optical mapping using arrayed, fluid fixed, DNA molecules. *Proc. Natl. Acad. Sci.* **95:** 8046–8051.

Lander, E.S. and M.S Waterman. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.

Lin, J., R. Qi, C. Aston, J. Jing, T.S. Anantharaman, B. Mishra, O. White, J.C. Venter, and D.C. Schwartz. 1998. Complete shotgun optical mapping of *Deinococcus radiodurans* and *Escherichia coli* K12 using genomic DNA molecules. Submitted.

Netzel, T.L., K. Nafisi, M. Zhao, J.R. Lenhard, and I. Johnson. 1995. Base-content dependence of emission enhancements, quantum yields, and lifetimes for cyanine dyes bound to double-strand DNA: Photophysical properties of monomeric and bichromophoric DNA stains. *J. Phys. Chem.* **99:** 17936–17947.

Schwartz, D.C. and C.R. Cantor. 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37:** 67–75.

Schwartz, D.C. and M. Koval. 1989. Conformational dynamics of individual DNA molecules during gel electrophoresis. *Nature* **338:** 520–522.

Schwartz, D.C., X. Li, L. Hernandez, S. Ramnarain, E. Huff, and Y. Wang. 1993. Ordered restriction maps of *Saccharomyces cerevisae* chromosomes constructed by optical mapping. *Science* **262:** 110–114.

Sinnis, P. and T.E. Wellems. 1988. Long-range restriction maps of *Plasmodium falciparum* chromosomes: Crossingover and size variation among geographically distant isolates. *Genomics* **3:** 287–295.

Trager, W. and J.B. Jensen. 1976. Human malaria parasites in continuous culture. *Science* **193:** 673–675.

Thompson, J.K. and A.F. Cowman 1997. A YAC contig and high resolution map of chromosome 3 from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **90:** 537–542.

Watanabe, J. and J. Inselburg. 1994. Establishing a physical map of chromosome No. 4 of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **65:** 189–199.

Received October 5, 1998; accepted in revised form December 15, 1998.

A shotgun optical map of the entire *Plasmodium falciparum* genome

Zhongwu Lai[1], Junping Jing[1], Christopher Aston[1], Virginia Clarke[1], Jennifer Apodaca[1], Eileen T. Dimalanta[1], Daniel J. Carucci[3], Malcolm J. Gardner[4], Bud Mishra[2], Thomas S. Anantharaman[2], Salvatore Paxia[2], Stephen L. Hoffman[3], J. Craig Venter[4], Edward J. Huff[1] & David C. Schwartz[1,5]

The unicellular parasite *Plasmodium falciparum* is the cause of human malaria, resulting in 1.7–2.5 million deaths each year[1]. To develop new means to treat or prevent malaria, the Malaria Genome Consortium was formed to sequence and annotate the entire 24.6-Mb genome[2]. The plan, already underway, is to sequence libraries created from chromosomal DNA separated by pulsed-field gel electrophoresis (PFGE). The AT-rich genome of *P. falciparum* presents problems in terms of reliable library construction and the relative paucity of dense physical markers or extensive genetic resources. To deal with these problems, we reasoned that a high-resolution, ordered restriction map covering the entire genome could serve as a scaffold for the alignment and verification of sequence contigs developed by members of the consortium. Thus optical mapping was advanced to use simply extracted, unfractionated genomic DNA as its principal substrate. Ordered restriction maps (*Bam*HI and *Nhe*I) derived from single molecules were assembled into 14 deep contigs corresponding to the molecular karyotype determined by PFGE (ref. 3).

Optical mapping is now a proven means for the construction of accurate, ordered restriction maps from ensembles of individual DNA molecules derived from a variety of clone types, including bacterial artificial chromosomes[4] (BACs), yeast artificial chromosomes[5] (YACs) and small insert clones[6]. We previously developed approaches for mapping clone DNA samples that relied on the analysis of large numbers of identical DNA molecules. Here, the challenge was to develop ways to generate restriction maps of a population of randomly sheared DNA molecules directly extracted from cells that were obviously non-identical. Problems to be solved included the development of techniques for mounting very large DNA molecules onto surfaces and new methods for accurately mapping individual molecules, which were uniquely represented within a population. Finally, new algorithms were necessary to assemble such maps into gap-free contigs covering all 14 chromosomes of the *P. falciparum* genome.

We developed an optical mapping approach, termed shotgun optical mapping, that used large (250–3,000 kb), randomly sheared genomic DNA molecules as the substrate for map construction (Fig. 1*a–e*). Random fragmentation of genomic DNA occurred naturally as a consequence of careful pipetting and other manipulations. Surface-mounted molecules were digested using *Bam*HI and *Nhe*I (refs 6–8). Because genomic DNA molecules frequently extended through multiple digital image fields, we developed an automated image acquisition system (GenCol) to overlap digital images with proper registration (Figs 1*c* and 2). Map construction techniques were altered to take into account local restriction endonuclease efficiencies (the rate of partial

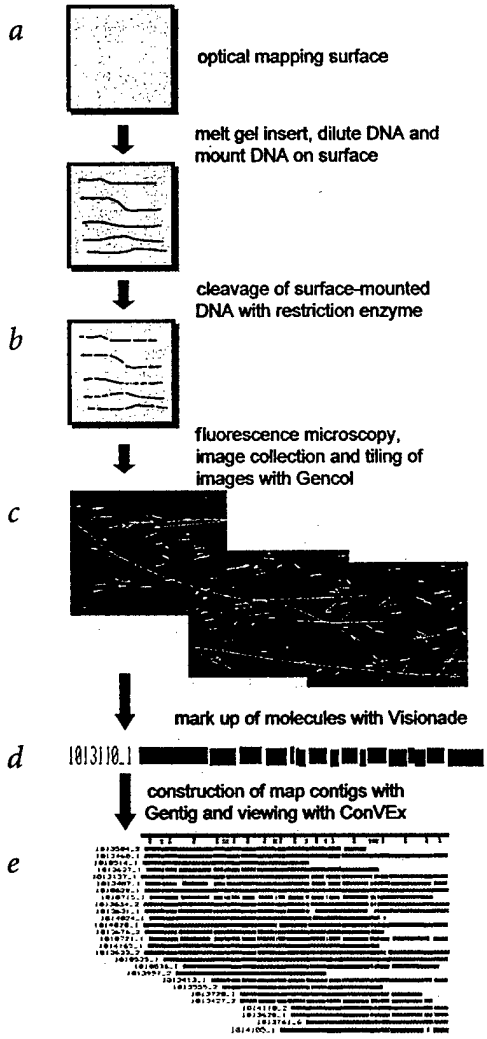

Fig. 1 Schematic of shotgun optical mapping approach. ***a***, Shotgun optical mapping used large (250–3,000 kb), randomly sheared genomic DNA molecules as the substrate for map construction. ***b***, Random fragmentation of genomic DNA occurred naturally as a consequence of careful pipetting and other manipulations. Surface-mounted molecules were digested using *Bam*HI and *Nhe*I (ref. 8). ***c***, Because genomic DNA molecules frequently extended through multiple image fields, an automated image acquisition system was developed (GenCol) and used to overlap images with proper registration. ***d***, Map construction techniques take into account local restriction endonuclease efficiencies (the rate of partial digestion) and the analysis of molecule populations that differed in composition and mass. ***e***, These steps were necessary to enable accurate construction of map contigs.

[1]W.M. Keck Laboratory for Biomolecular Imaging, Department of Chemistry and [2]Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, Department of Chemistry, New York, New York, USA. [3]Malaria Program, Naval Medical Research Center and [4]The Institute for Genomic Research, Rockville, Maryland, USA. [5]Present address: University of Wisconsin-Madison, Departments of Chemistry and Genetics, UW Biotechnology Center, Madison, Wisconsin, USA. Correspondence should be addressed to D.C.S. (e-mail: schwad01@mcrcr.med.nyu.edu).

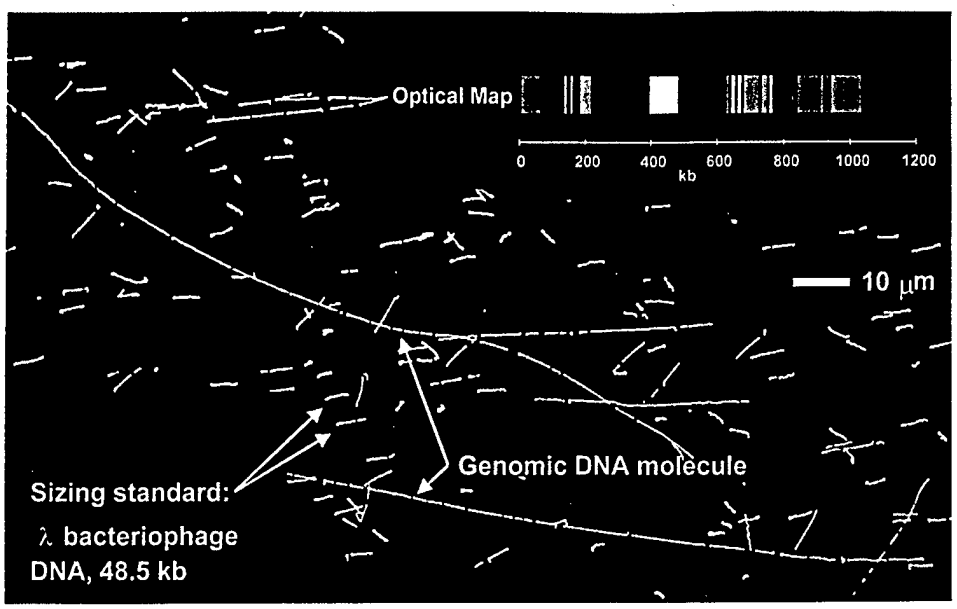

Fig. 2 Digital fluorescence micrograph and map of a typical genomic DNA molecule. A *P. falciparum* molecule digested with *Nhe*I is shown with its corresponding optical map. Comparison with the consensus optical map shows this molecule to be an intact chromosome 3. Image composed by tiling a series of 63× (objective power) images using GenCol. Co-mounted λ bacteriophage DNA is used as a sizing standard and to estimate cutting efficiencies.

digestion) and the analysis of molecule populations that differed in composition and mass. These steps were necessary to enable accurate construction of map contigs.

Previous map construction techniques using cloned DNA molecules[5,6,9] determined restriction-fragment mass on the basis of relative measures of integrated fluorescence intensities or apparent lengths. Thus, fragment masses were reported as a fraction of the total clone size (1.0), and later converted to kilobases by independent measure of clone masses (that is, cloning vector sequence[10]). Additionally, maps derived from ensembles of identical molecules were averaged to construct final maps. In contrast, here, we independently sized restriction fragments in genomic shotgun optical mapping using λ bacteriophage DNA that was co-mounted and digested in parallel (Fig. 2). These molecules were also used to locally monitor the restriction digestion efficiency, and to infer the extent of digestion on a per molecule (genomic) basis. Cutting efficiencies were in excess of 80%. This assessment provided a critical set of parameters for the contig assembly program, 'Gentig'[8,11,12], to reliably overlap maps derived from individual DNA molecules.

Gentig assembled maps into a number of deep contigs, but did not assign every single-molecule map to a contig. The program assembled contigs using 50% of the available molecules, which corresponded to 70% of the total mass of the molecules. In other words, the program was better able to construct contigs from the longer single-molecule maps. Finishing work using spreadsheets assembled the data into 14 contigs corresponding to the PFGE-generated molecular karyotype, with a total genome size of 24.16 Mb (Table 1). *Bam*HI and *Nhe*I maps had an average fragment size of 30.6 kb and 30.1 kb, respectively. We constructed consensus maps (Fig. 3) by simple averaging of aligned restriction-fragment masses (typically 6–26 fragments) derived from overlapping DNA molecules. Overall, chromosome sizes were largely consistent with PFGE results, with the total optical genome size being approximately 7% smaller, indicating that no previously uncharacterized nuclear component was found.

We previously constructed a high-resolution optical map of *P. falciparum* chromosome 2 (ref. 7). The starting material was a PFGE gel slice containing fractionated chromosome 2 DNA. We now constructed a whole-genome optical map using total, unfractionated genomic DNA as the starting material and resolved all 14 chromosomes, including electrophoretically unseparable ones (chromosomes 5–9, termed the 'blob'), at the level of data (optical map contigs) rather than as physical entities (that is, gel bands).

Table 1 • *P. falciparum* whole-genome optical mapping

Chr.	PFGE (Mb)	*Nhe*I (Mb)	*Bam*HI (Mb)	Ave. (Mb)	Diff. (Mb)	Linkage/ confirmation	Orientation
1	0.65/0.65*	0.684	0.668	0.676	0.016	1,3	+
2	1.0/0.947*	0.958	1.037	0.997	0.079	1,3	+
3	1.2/1.060*	1.084	1.096	1.090	0.012	1,2	+
4	1.4	1.311	1.306	1.309	0.005	1	
5	1.6	1.331	1.337	1.334	0.006		
6	1.6	1.395	1.373	1.384	0.022		
7	1.7	1.494	1.444	1.469	0.050		
8	1.7	1.495	1.504	1.499	0.009		
9	1.8	1.600	1.595	1.598	0.005		
10	2.1	1.808	1.688	1.748	0.120	1	
11	2.3	2.097	2.089	2.093	0.008	1	
12	2.4	2.478	2.361	2.419	0.117	1	
13	3.2	3.172	3.022	3.097	0.150	2	+
14	3.4	3.436	3.404	3.420	0.032	1,3	+
Total	26.05	24.341	23.974	24.157	0.367		

*Size from sequencing. Linkage/confirmation was obtained as follows: by mapping PFGE-purified chromosomal material (1); by mapping chromosome-specific YACs (2); or by sequence information (3). +, *Bam*HI and *Nhe*I maps have been oriented. Chr., chromosome; Ave., average size; Diff., difference between *Bam*HI and *Nhe*I maps.

Fig. 3 High-resolution optical mapping of the *P. falciparum* genome using *Nhe*I and *Bam*HI. We mapped 944 molecules with *Nhe*I; the average molecule length was 588 kb, corresponding to 23× coverage. We mapped 1,116 molecules with *Bam*HI; the average molecule length was 666 kb, corresponding to 31× coverage. ***a***, Gap-free, consensus *Nhe*I and *Bam*HI maps were generated across all 14 *P. falciparum* chromosomes using the map contig assembly program Gentig. ***b,c***, *Nhe*I and *Bam*HI map alignments determined by Gentig, displayed by ConVEx. Fragment sizes of consensus maps (blue lines) shown in (*a*) were determined from the alignment and averaging of maps derived from 6–26 underlying individual molecules (green lines), 230–2,716 kb. ***d***, Enlargement of contig for chromosome 3 (*Nhe*I) shown in ConVEx displays maps (green) scaled to the consensus map (blue). These data can be accessed at http://carbon.biotech.wisc.edu/plasmodium. Bar lengths reflect measured fragment sizes. Fragments that overlap are shaded.

To assess errors produced by shotgun optical mapping, we compared optical restriction maps for chromosome 2 with restriction maps generated *in silico* using a previously assembled sequence[13]. We found good correspondence between the two maps. The sequence shows chromosome 2 to be 947 kb versus 958 kb by optical mapping with *Nhe*I and 1,037 kb with *Bam*HI. Only one 600-bp *Bam*HI fragment was missing in the entire genome optical map. The *Nhe*I optical map included all fragments above 400 bp predicted from sequence. The average absolute relative error in sizing fragments was 4.6% for *Nhe*I and 5.0% for *Bam*HI. Likewise, similar errors for chromosome 3 were determined by comparing optical maps with sequence data (*Nhe*I, 4.4%; *Bam*HI, 4.1%; total optical size versus sequence, *Nhe*I, 1,084 kb; *Bam*HI, 1,147 kb; versus 1,060 kb; D. Lawson, pers. comm.). These sizing errors were similar to those associated with PFGE.

Some large *Nhe*I and *Bam*HI fragments were noticeable at the telomeric ends. A telomere of one of the 'blob' chromosomes (chromosome 7) is composed of three consecutive 6-kb *Bam*HI fragments. Optical mapping can estimate numbers of repetitive regions if the repeats contain recognition sites for the endonuclease used. Subtelomeric regions in *P. falciparum*, however, are characterized by 21-bp tandem repeats[14], which are too small to be detected by optical mapping.

We used several approaches to verify and to link our optical maps with the PFGE molecular karyotypes, which number chromosomes according to mobility. Chromosomes that were identified and the orientations of *Bam*HI and *Nhe*I maps are shown

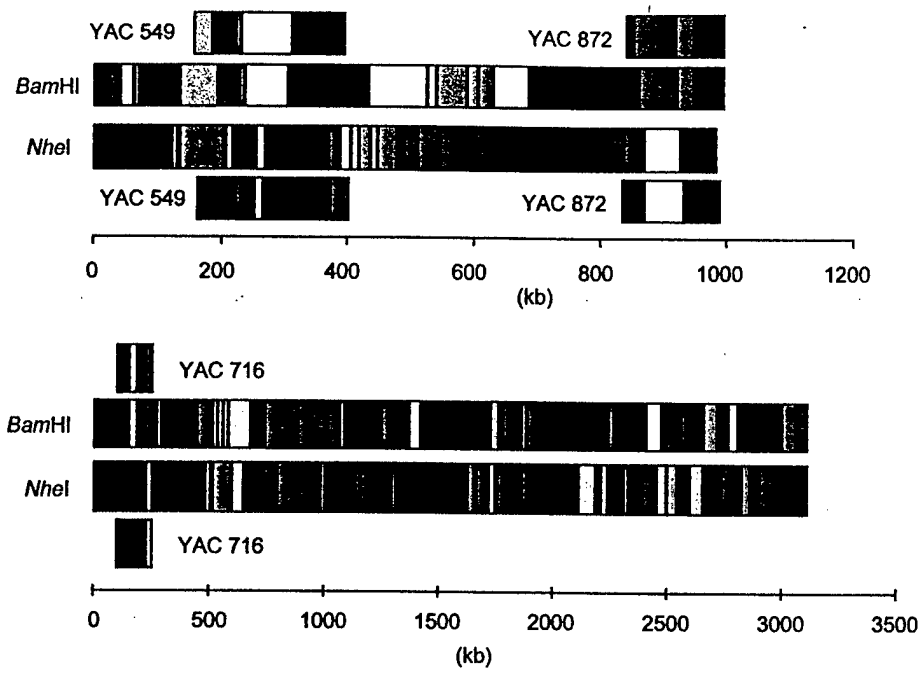

Fig. 4 Identification of chromosomes and alignment of *Nhe*I and *Bam*HI maps by mapping chromosome-specific YAC clones. Chromosome 3 and 13-specific YAC maps were aligned with the optical maps and the two enzyme maps were then oriented and linked. Each YAC is ~150 kb.

(Table 1). We confirmed chromosomal identities of some optical maps by optical mapping of PFGE-purified chromosomal DNA (ref. 7) with *Nhe*I or *Bam*HI. Here, most maps formed a contig, which aligned with a specific consensus map. Despite the fact that the largest and smallest *P. falciparum* chromosomes are resolved by PFGE, the gel slices contained DNA molecules from other chromosomes. There was, however, a sufficiently large population of molecules that formed a contig with a particular chromosome (>50%) to be able to identify it as being the chromosome predicted from the molecular karyotype. When many chromosomes are similar in size, such as chromosomes 5–9, there are many possible orientations of the maps, thus this approach was not viable. Chromosome-specific YAC clones were also optically mapped for further confirmation of chromosomal orientation and linkage. We aligned the resulting maps with a specific contig in the consensus maps (Fig. 4). YAC clones were not available for those chromosomes in the 'blob', so we were unable to identify or link these optical maps. As such, we have assigned numbers to these chromosomes according to their optically determined masses (Table 1). Maps can also be linked together by a series of double digestions, by the use of available sequence information or by Southern blot using chromosome-specific probes.

Because unicellular parasites have relatively small chromosomes that do not visibly condense, PFGE has provided a means by which chromosomal entities can be physically mapped and studied at the molecular level[15–17]. In fact, PFGE separations are currently providing the very material that the international malaria consortium is using to create chromosomal-specific libraries for large-scale sequencing efforts (http://www-ermm.cbcu.cam.ac.uk/dcn/txt001dcn.htm). Unfortunately, parasites such as *P. falciparum* can have karyotypically complex genomes, which confound PFGE analysis by displaying similarly sized chromosomes. Furthermore, very large or circular chromosomes are difficult to physically identify or characterize[18]. Although the shotgun sequencing of entire microorganism genomes[19,20] has obviated physical mapping to some extent, high-quality, finished sequence remains laborious to generate.

Many issues regarding the efficient sequencing of lower eukaryotes remain to be fully resolved, especially when available map resources are minimal. In the case of *Saccharomyces cerevisiae*, the entire genome was sequenced by a large consortium of laboratories on a per chromosome basis[21]. Their tasks were facilitated by the availability of extensive physical and genetic maps, plus an assortment of well-characterized libraries. These substantial genome resources provided ample means for the needed sequence verification efforts, and aids for the sequence-assembly process. In a similar, though much less distributive fashion, the *Caenorhabditis elegans* genome was recently completely sequenced[22]. Given the rapid pace of electrophoretic sequencing technology[23,24] and the accumulation of resources in sequence acquisition and analysis, new ways to efficiently sequence lower eukaryotes, particularly those implicated in human disease, must be developed to optimally leverage map resources created by optical mapping.

The optical maps presented here have been used by members of the consortium[13,25] as scaffolds to verify and facilitate sequence assemblies. In general, the maps were integrated into the sequence assembly process, in much the same way as any other physical maps. In particular, our maps have provided reliable landmarks for sequence assembly where traditional maps are somewhat sparse. Compared with sequence-tagged site (STS) or EST maps, in which landmark order is known but physical distance is approximate, optical restriction maps are constructed from landmarks (restriction sites) that are precisely characterized by physical distance. Another advantage is the speed of map construction: the maps presented here required only 4–6 months to generate. Given these and other advantages, future work will center on the algorithmic integration of high-resolution optical maps with primary sequence reads to more fully automate the sequence assembly and verification process. Finally, we plan to use optical mapping as the basis for developing of new ways to study genomic variations that fall between, or outside of, the capabilities of sequence-based approaches and cytogenetic observation.

Methods

Parasite preparation. We cultivated *P. falciparum* (clone 3D7) in erythrocytes using standard techniques[26]. Possible alterations of the genome that can occur in continuous culture[27] were minimized by keeping parasite aliquots frozen in liquid N_2 until needed. We then cultivated parasites only as long as necessary and prepared agarose-embedded parasites as described[7].

Mounting and digestion of DNA on optical mapping surfaces. We prepared derivatized glass optical mapping surfaces as described[7,28]. We diluted genomic DNA in TE buffer containing a sizing standard (λ bacteriophage DNA, 50 ng/ml), which was co-mounted with the genomic DNA by spreading the sample into the space between the surface and a microscope slide. DNA molecules were digested with *Nhe*I or *Bam*HI (ref. 8). λ bacteriophage DNA (48.5 kb; New England Biolabs) is cut once by *Nhe*I. λ DASH II bacteriophage DNA (41.9 kb; Stratagene) is cut twice by *Bam*HI. Therefore, we also used standards to identify regions on the surface where the digestion efficiency exceeded 70%. We stained DNA with YOYO-1 homodimer (Molecular Probes), before fluorescence microscopy. *P. falciparum* DNA has an AT content of 80–85%, and λ bacteriophage DNA has an AT content of 50%. The YOYO-1 fluorochrome used for DNA staining preferentially intercalates between GC pairs with increased emission quantum yield[29]. We therefore applied a correction factor to each fragment size to correct for this variation in fluorochrome incorporation.

Image acquisition, processing and map construction. We collected digital images of DNA molecules with a cooled charge coupled device (CCD) camera (Princeton Instruments) using Optical Map Maker (OMM) software as described[6]. Because genomic DNA molecules span multiple microscope image fields, we developed 'GenCol', an image acquisition and management software that was used to automatically collect and overlap consecutive CCD images with proper pixel registration. GenCol used a precise fitting routine, and the resulting 'super-images' covered the entire length of single DNA molecules, spanning several microscope fields. Restriction fragments were marked up with 'Visionade'[28], a semi-automatic visualization/editing program, which was run on super-images. Files created from marked-up images of molecules were then sent to map construction software, which automatically determined the restriction fragment masses, characterized internal DNA standard molecules and produced finished maps from single genomic molecules. The integrated fluorescence intensities of λ bacteriophage DNA standards, co-mounted with the genomic molecules, were used to measure the size of the *P. falciparum* restriction fragments on a per image basis. Cutting efficiencies (on a per image basis) were determined from scoring cut sites on sizing standard molecules contained in the same field as the genomic DNA molecules. Knowledge of endonuclease cutting efficiencies was critical for accurate contig construction.

Contig assembly by Gentig. Sophisticated statistical methods are used to overcome errors associated with partial digestion and mass determination[11,12]. Gentig finds overlapped molecules and assembles them into contigs. It computes contigs of genomic maps using a heuristic algorithm for finding the best scoring set of contigs (overlapping maps), because finding the optimal placement is in general computationally too expensive. The entire *P. falciparum* genome data set can be assembled into contigs in ~20 min. Gentig assembled consensus maps for each chromosome by averaging the fragment sizes from the individual maps underlying the contigs.

Contig viewing and editing by 'ConVEx'. We viewed contigs using 'ConVEx' (contig visualization and exploration tool). ConVEx is a multi-scale zoomable interface for visualization and exploration of large, high-resolution contiged restriction maps. Users can examine the consensus maps together with the raw uncorrected data. ConVEx also has a 'lens' mechanism that provides annotation and editing features, allowing communication of features such as STS markers, and even the underlying sequence reads.

Chromosome isolation by PFGE. The genome of *P. falciparum* is ~25 Mb, consisting of 14 chromosomes ranging from 0.6 to 3.5 Mb (ref. 28). PFGE resolves most of the *P. falciparum* chromosomes, except 5–9, which are of similar sizes and co-migrate. PFGE-purified chromosomal DNA was prepared as described[8] and used as a substrate for optical mapping.

YAC isolation and mapping. We cultured yeast cells in AHC media and prepared agarose-embedded cells using standard methods[3]. We purified YAC DNA using PFGE (POE apparatus, 1% gel in 0.5×TBE, pulse time 3 s, 5 s; switch time 32 s; 150 volts for 24 h; ref. 30). Optical maps of YAC clones were prepared with *Nhe*I and *Bam*HI as described above.

Acknowledgements

We thank D. Lawson and T. Wellems for clones and other valuable reagents. This work was supported by the Burroughs Wellcome Fund, NIH, and the Naval Medical Research and Development Command work unit STEP C611102A0101BCX. Additional support came from NCHGR (2 RO1 HG00225-01-09) and NCI1(RO1CA 79063-1).

Received 26 May; accepted 20 September 1999

1. World Health Organization. World malaria situation in 1994. Part I. Population at risk. *Wkly Epidemiol. Rec.* **72**, 269–274 (1997).
2. Wirth, D. Malaria: a 21st century solution for an ancient disease. *Nature Med.* **4**, 1360–1362 (1998).
3. Schwartz, D.C. & Cantor, C.R. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67–75 (1984).
4. Cai, W., Aburatani, H., Housman, D., Wang, Y. & Schwartz, D.C. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc. Natl Acad. Sci. USA* **92**, 5164–5168 (1995).
5. Cai, W. *et al.* High resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proc. Natl Acad. Sci. USA* **95**, 3390–3395 (1998).
6. Jing, J. *et al.* Automated high resolution optical mapping using arrayed, fluid fixed, DNA molecules. *Proc. Natl Acad. Sci. USA* **95**, 8046–8051 (1998).
7. Jing, J. *et al.* Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Res.* **9**, 175–181 (1999).
8. Lin, J. *et al.* Whole genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* **285**, 1558–1562 (1999).
9. Schwartz, D.C. & Samad, A. Optical mapping approaches to molecular genomics. *Curr. Opin. Biotechnol.* **8**, 70–74 (1997).
10. Meng, X., Benson, K., Chada, K., Huff, E.J. & Schwartz, D.C. Optical mapping of lambda bacteriophage clones using restriction endonucleases. *Nature Genet.* **9**, 432–438 (1995).
11. Anantharaman, T.S., Mishra, B. & Schwartz, D.C. Genomics via Optical Mapping III: contiging genomic DNA and variations. in *Courant Technical Report 760* (Courant Institute, New York University, New York, 1998).
12. Anantharaman, T.S., Mishra, B. & Schwartz, D.C. Genomics via Optical Mapping III: contiging genomic DNA and variations. *The Seventh International Conference on Intelligent Systems for Molecular Biology* **7**, 18–27 (1999).
13. Gardner, M.J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132 (1998).
14. Pace, T., Ponzi, M., Scotti, R. & Frontali, C. Structure and superstructure of *Plasmodium falciparum* subtelomeric regions. *Mol. Biochem. Parasitol.* **69**, 257–268 (1995).
15. Van der Ploeg, L.H.T., Schwartz, D.C., Cantor, C.R. & Borst, P. Antigenic variation in *Trypanosoma brucei* analyzed by electrophoretic separation of chromosome sized DNA molecules. *Cell* **37**, 77–84 (1984).
16. Spithill, T.W. & Samaras, N. The molecular karyotype of *Leishmania major* and mapping of α and β tubulin gene families to multiple unlinked chromosomal loci. *Nucleic Acid Res.* **13**, 4155–4169 (1985).
17. Ahamada, S., Wery, M. & Hamers, R. Rodent malaria parasites: molecular karyotypes characterize species, subspecies and lines. *Parasite* **1**, 31–38 (1994).
18. Moritz, K.B. & Roth, G.E. Complexity of germline and somatic DNA in *Ascaris*. *Nature* **259**, 55–57 (1976).
19. Fleischmann, R.D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
20. Fraser, C.M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
21. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–567 (1996).
22. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
23. Mullikin, J.C. & McMurray, A.A. Sequencing the genome, fast. *Science* **283**, 1867–1868 (1999).
24. Venter, J.C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540–1542 (1998).
25. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538 (1999).
26. Trager, W. & Jensen, J.B. Human malaria parasites in continuous culture. *Science* **193**, 673–675 (1976).
27. Corcoran, L.M., Forsyth, K.P., Bianco, A.E., Brown, G.V. & Kemp, D.J. Chromosome size polymorphism in *Plasmodium falciparum* can involve deletions and are frequent in nature parasite populations. *Cell* **44**, 87–95 (1986).
28. Aston, C., Hiort, C. & Schwartz, D.C. Optical mapping: an approach for fine mapping. *Methods Enzymol.* **303**, 55–73 (1999).
29. Netzel, T.L., Nafisi, K., Zhao, M., Lenhard, J.R. & Johnson, I. Base-content dependence of emission enhancements, quantum yields, and lifetimes for cyanine dyes bound to double-strand DNA: photophysical properties of monomeric and bichromophoric DNA stains. *J. Phys. Chem.* **99**, 17936–17947 (1995).
30. Schwartz, D.C. & Koval, M. Conformational dynamics of individual DNA molecules during gel electrophoresis. *Nature* **338**, 520–522 (1989).

optical map

BioWorld® Today

Thursday
November 4, 1999

THE DAILY BIOTECHNOLOGY NEWSPAPER

Volume 10, No. 211
Page 1 of 7

ICOS Receives $15M Milestone But Also Has Two Failed Trials

By Mary Welch
Staff Writer

ICOS Corp. received a $15 million milestone for starting a Phase III study of its treatment of erectile dysfunction, but the company also said that two other drugs in Phase II trials failed to achieve statistical significance in their endpoints.

ICOS, of Bothell, Wash., and partner Eli Lilly & Co., of Indianapolis, will take IC351 into Phase III trials to test the phosphodiesterase type 5 (PDE5) inhibitor as an oral treatment for sexual dysfunction in men and women.

"We have quite a number of compounds for a number of indications in the clinic," said Gary Peterman, ICOS' senior director of therapeutic development. "We understand that not all of them would be a success in all indications. We saw the result of that today. But we are confident of our long-term strategy and very pleased with what's happening with IC351 and the $15 million milestone."

See ICOS, Page 4

BioCryst Brings In $50.5M Through Public Offering

By Lisa Seachrist
Washington Editor

With competition increasing in the market for flu drugs, small-molecule specialist BioCryst Pharmaceuticals Inc. raised $50.5 million in a public offering to fund development of its flu pill and other clinical and preclinical programs.

The Birmingham, Ala.-based company sold 2 million shares at a price of $25.25 each, exceeding its expected offering revenues of $45.5 million based on a share price of $25. The money will be used to advance its lead flu drug, RWJ-270201, formerly known as BCX-1812, as well as development programs in T-cell related diseases and purine nucleoside phosphorylase (PNP) inhibitors.

The underwriters for the offering – Salomon Smith Barney Inc., of New York; Hambrecht & Quist LLC, of San Francisco; and Raymond James & Associates Inc., of New York – were granted an option to purchase an additional 300,000

See BioCryst, Page 3

Consortium Takes On Plasmodium Falciparum

Sequencing of Malarial Parasite Genome Gets Cutting-Edge Boost From Optical Mapping Technique

By David N. Leff
Science Editor

When a lone serial killer goes to ground, law-enforcement elements organize a search party made up of local cops from the crime scene, county sheriffs, state constabulary, U.S. marshals and the FBI. Once they hunt down and catch the suspect, forensic DNA tests help nail down the culprit's identity, and evidence linking him to his victims.

Now a multinational posse of genomicists is hot on the trail of the world's largest-scale killer, most of whose victims are children. Their quarry is the mosquito-borne parasite, *Plasmodium falciparum*, which takes the lives of some 2 million people a year, most of them in the world's tropical areas.

See Genome, Page 5

Versicor Raises $40 Million To Develop Lead Compounds

By Mary Welch
Staff Writer

Versicor Inc. closed a $40 million round of private equity financing aimed at funding its trials of LY30336, an antifungal class of drugs, and BI 397, its second-generation glycopeptide.

"This is significant funding," said George Horner, president and CEO of the Fremont, Calif.-based company. "We wanted to raise $36 million, so this gives us a lot of flexibility. It will take us through until we are able to go public."

Horner declined to speculate when the company, which spun off from Sepracor Inc. in 1995, would enter the public market.

Horner also refused to say how many shares of stock exchanged hands or what percentage of the company's equity was involved. "We're a private company and I just

See Versicor, Page 3

Genome

Continued from Page 1

While other task forces have been striving for decades to develop anti-malarial drugs and vaccines, the U.S.-UK Malaria Genome Consortium came into existence a few years ago to sequence *P. falciparum*'s entire 24.6 megabase genome.

"The consortium," said genomicist David Schwartz, at the University of Wisconsin-Madison, "is funded by the U.S. National Institutes of Health, and Britain's Wellcome Trust plus the Burroughs-Wellcome Fund. The major sequencing efforts," he added, include labs from Stanford University; TIGR (The Institute for Genomic Research, in Rockville, Md.); and the UK's Sanger Institute. "Those people have divvied up the parasite's 14 chromosomes," Schwartz observed.

"The consortium's goals," he pointed out, "are first completely analyzing the genome, in terms of getting its sequence. That should take about a year from now, maybe a little bit longer. Then, secondly, making sense of that sequence."

Schwartz himself is making some of that sense at a pre-sequencing stage of the consortium's progress. He is senior author of an article in the November 1999 issue of *Nature Genetics*, titled, "A shotgun optical map of the entire *Plasmodium falciparum* genome."

"Optical mapping," Schwartz explained, "is a new technology, which colleagues and I invented and patented 10 or 11 years ago. It maps an organism's entire genome from single DNA molecules, provides reliable landmarks, and could ratchet up the race to decipher complete genomes – from food crops to human beings.

"One can picture optical mapping as an entire map of the United States," he suggested, "whereas conventional genome sequencing would be thousands of detailed maps of every city in the nation. Optical mapping data works in concert with high-resolution DNA sequence data, linking both together in a complete and seamless description of a genome."

Consortium laboratories, Schwartz told *BioWorld Today*, "are already incorporating our optical scanning of *P. falciparum* into their total-sequencing modus operandi. They are relying on us to do the optical mapping, and they do the sequencing."

Parasite's Map – From Bottle To Data

The 15 co-authors of his *Nature Genetics* report reflect the assembly-line procedure of optically mapping the parasite's genome. "At the U.S. Naval Medical Research Center's malaria program," Schwartz recounted, "Daniel Carucci and his colleagues were able to grow the single-cell parasite, put it into a bottle, then simply extract DNA from these bottled pathogens."

"Then he sent the DNA to us – just long strands, millions of bases in length. We simply pinned them down on glass surfaces, to which they stick through electrostatic forces. It's very much like rubbing a balloon on a wool sweater; then you're able to stick it to a wall. Basically, the DNA molecules stick to our glass surface, and elongate.

"Next," Schwartz went on, "we took two restriction enzymes, and cut the DNA strands. Wherever an enzyme recognizes its cognate site, it cuts the molecule. And we can see that it cuts because that's where a gap forms – visible enough so that we can see it through a light microscope. We have software that automatically goes in there and finds the cleaved fragments, and measures their mass, their size, according to how much fluorescence is associated with each fragment.

"When you have an ordered restriction map," he went on, "a single molecule that's been cut, it generates a series of daughter fragments, which constitute a single ordered restriction map. Then software puts together many such maps that have overlaps of commonality with one another, at least in part – and that did it. We then had a physical map of an entire *P. falciparum* genome, generated without clones or PCR or electrophoresis."

For purposes of genome sequencing, Schwartz pointed out, "these maps serve as a scaffold to tell you very concisely how to align small snippets of sequence with a whole chromosome. Also, to know if your sequence is correct, this is a way of error-checking it."

Anti-Malarial Drugs, Vaccines, Therapies?

"The fact that optical mapping can facilitate sequencing," Schwartz pointed out, "and be sure about it, provides the stuff that people are going to be looking at to develop new anti-malarial therapies, new vaccines, new drugs and so on. By comparing maps of hundreds of individual human genomes, for example," he added, "scientists could pinpoint the origin of genetic diseases, understand the complexities of trait inheritance, examine the process behind DNA repair. This is like the Periodic Table."

Having wrapped up *P. falciparum* – which took them five months – he and his co-authors are now tackling the genome of *Trypanosoma brucei*, the pathogen of African sleeping sickness. And he has just received a grant to take on the genome of rice, the world's No. 1 food crop.

"What we've been in the process of doing for the past three or four years," Schwartz said, "is trying to harden the optical mapping system so that it will have a very very high throughput, and be very cheap to do. Right now, we're looking for industrial partners to do that, because this sort of development work doesn't go that well in a university environment. Originally we got a lot of funding from Chiron and Novartis. So now," he concluded, "we're thinking about trying to put together our own company." ■

Parassitologia 41: 69-75, 1999

The malaria genome sequencing project: complete sequence of *Plasmodium falciparum* chromosome 2

M.J. Gardner[1], H. Tettelin[1], D.J. Carucci[2], L.M. Cummings[1], H.O. Smith[1], C.M. Fraser[1], J.C. Venter[1], S.L. Hoffman[2]

[1] *The Institute for Genomic Research;* [2] *Malaria Program, Naval Medical Research Center, Rockville, MD, USA.*

Abstract. An international consortium has been formed to sequence the entire genome of the human malaria parasite *Plasmodium falciparum*. We sequenced chromosome 2 of clone 3D7 using a shotgun sequencing strategy. Chromosome 2 is 947 kb in length, has a base composition of 80.2% A+T, and contains 210 predicted genes. In comparison to the *Saccharomyces cerevisiae* genome, chromosome 2 has a lower gene density, a greater proportion of genes containing introns, and nearly twice as many proteins containing predicted non-globular domains. A group of putative surface proteins was identified, rifins, which are encoded by a gene family comprising up to 7% of the protein-encoding genes in the genome. The rifins exhibit considerable sequence diversity and may play an important role in antigenic variation. Sixteen genes encoded on chromosome 2 showed signs of a plastid or mitochondrial origin, including several genes involved in fatty acid biosynthesis. Completion of the chromosome 2 sequence demonstrated that the A+T-rich genome of *P. falciparum* can be sequenced by the shotgun approach. Within 2-3 years, the sequence of almost all *P. falciparum* genes will have been determined, paving the way for genetic, biochemical, and immunological research aimed at developing new drugs and vaccines against malaria.

Key words: *Plasmodium falciparum*, malaria, chromosome 2, rifins, genomics, malaria genome sequencing project.

In 1995, the first complete genome sequence of a free-living organism, *Haemophilus influenzae*, was published (Fleischmann *et al.*, 1995). The publication of the *H. influenzae* genome sequence marked a turning point in biology. As noted by Bloom, it heralded a post-genomics era of microbe biology when the complete genomes of most human pathogens would have been sequenced, providing a vast database of sequence information that would enable researchers to focus on studies of the biology and pathogenicity of these organisms (Bloom, 1995). This research in turn would lead to the development of new drugs and vaccines to treat and prevent diseases caused by these pathogens, and would be especially useful for research on organisms difficult to grow. Since then, there has been a flurry of effort to sequence the genomes of other pathogens, and the genomes of organisms that cause diseases such as syphylis (*Treponema pallidum*), ulcers (*Helicobacter pylori*), Lyme disease (*Borrelia burgdorferi*), tuberculosis (*Mycobacterium tuberculosis*), and trachoma (*Chlamydia trachomatis*) have been completed (Fraser *et al.*, 1997, 1998; Tomb *et al.*, 1997; Cole *et al.*, 1998; Stephens *et al.*, 1998). The genomes of several microbes of environmental importance have also been sequenced, as has the genome of the yeast *Saccharomyces cerevisiae* (see <http://www.tigr.org/tdb/mdb/mdb.html> for a complete listing of microbial genomes that have been sequenced or that are in progress). There is no evidence, so far, that the pace of sequencing has slackened, and that more than 60 microbial genomes is currently underway.

The completion of the first few microbial genomes caused several groups to contemplate the sequencing of the *Plasmodium falciparum* genome. It was realized that determination of the complete *P. falciparum* genome sequence would be of great value to malariologists given the difficulty of studying this organism in the laboratory, with large parts of the life cycle being difficult, expensive, or impossible to maintain in the laboratory. Furthermore, techniques such as DNA microarrays and transfection had been developed, providing researchers with new tools to study the expression and function of genes and gene products in malaria parasites. Several groups initiated pilot sequencing projects, and an international consortium including malaria researchers, genome laboratories, bioinformatics centers, and funding agencies was formed to coordinate the project, facilitate collaboration, and ensure that the data would be provided to the scientific community in a timely and useful manner (Hoffman *et al.*, 1997). The consortium met every 6 months during the start-up phase of the project and continues to meet regularly as the work proceeds.

At the time the *P. falciparum* project was started,

Invited contribution to the Malariology Centenary Conference "The malaria challenge after one hundred years of malariology" held in Rome at the Accademia Nazionale dei Lincei, 16-19 November 1998.

Correspondence: Dr Malcolm J. Gardner, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, Tel ++1 301 8383519, Fax ++1 301 8380208, e-mail: gardner@tigr.org

several prokaryotic and archaeal genomes had been finished, and sequencing of the genomes of yeast and *Caenorhabditis elegans* were nearing completion. Two strategies had been used in these projects. The clone-by-clone method, used to sequence the *Escherichia coli* and *S. cerevisiae* genomes, for example, involved sequencing of large-insert clones from cosmid, lambda, and YAC libraries (Blattner *et al.*, 1997). The clones sequenced were selected after the construction of a physical map, which provided a tiling path of overlapping clones spanning the genome. The other method, pioneered at TIGR, was the whole genome shotgun method, which used a genomic library of sheared 1-2 kb fragments prepared in a plasmid vector (Fleischmann *et al.*, 1995). Thousands of randomly selected small insert clones were picked and sequenced, and custom fragment assembly software was used to assemble the overlapping fragments into a contiguous sequence. This method proved to be very efficient in that construction of a physical map was not required prior to sequencing. However, very robust software for fragment assembly had to be developed that was able to handle many thousands of individual sequence reads and also deal with the repetitive sequences present in bacterial genomes. In addition, relational databases and software were developed to manage the gap closure, finishing, and annotation processes.

Sequencing of the *P. falciparum* genome raised some formidable technical challenges, however. At ~28 Mb, the *P. falciparum* genome was almost 20-fold larger than the *H. influenzae* genome and seemed too large to tackle by the whole genome shotgun method because of the computational requirements of the assembly process. Closure of the many gaps that would have remained after the initial assembly would also have been difficult with such a large genome and few sequence markers to guide the closure process. On the other hand, the clone-by-clone approach was ruled out because large-insert (>20 kb) genomic libraries of very AT-rich *P. falciparum* DNA in plasmid, lambda, and cosmid vectors that could be used for sequencing were not available. Although large-insert yeast artificial chromosome (YAC) libraries of *P. falciparum* (Foster and Thompson, 1995) had been constructed which appeared to be stable, YACs are not very well suited to high-throughput sequencing projects. Consequently, a new approach was adopted in which individual chromosomes were resolved on pulsed-field gels and used to prepare chromosome-specific shotgun libraries in plasmid and M13 vectors. Randomly-selected clones were then sequenced and assembled in the same way as for a whole-genome shotgun project. Some laboratories also performed low-coverage sequencing of shotgun libraries prepared from YACs previously mapped on the chromosomes (Foster and Thompson, 1995); the YAC shotgun sequences helped to group sequences from the same part of the chromosome and assisted in gap closure. Adoption of the chromosome-by-chromosome shotgun strategy allowed the sequencing effort to be distributed among the different sequencing centers.

Three groups are involved in the sequencing effort: TIGR and the Malaria Program of the US Naval Medical Research Center (NMRC); the Sanger Centre in the UK; and Stanford University. The current status of the project (as of July 1999) is summarized in Table 1. Once the problems that had been encountered in library construction, sequencing, assembly and gap closure were solved, all 3 groups began to make rapid progress. The complete sequence of chromosome 2 (0.95 Mb) was recently published by the TIGR/NMRC group (Gardner *et al.*, 1998), and the Sanger Center has virtually finished chromosome 3 (1.1 Mb). Work on the other chromosomes is well underway. The chromosome 2 sequence was submitted to GenBank and the sequence and annotation is available at TIGR's web site and at the NCBI (Table 1). Preliminary unedited sequence data is also available for downloading, browsing or searching on web sites maintained at each laboratory.

Table 1. Chromosome assignments and current status of the Malaria Genome Sequencing Project. [a] Estimated chromosome sizes for *P. falciparum* clone 3D7 were taken from Dame *et al.* (1996) or from the sequence data. [b] NIAID, National Institute for Allergy and Infectious Diseases; DoD, US Department of Defense; BWF, Burroughs Wellcome Fund. [c] Complete annotation (chromosome 2) or preliminary data can be viewed at web sites maintained by the sequencing centers: TIGR/NMRC <http://www.tigr.org/tdb/mdb/pfdb/pfdb.html>; the Sanger Centre <http://www.sanger.ac.uk/Projects/P_falciparum/>; Stanford University <http://baggage.stanford.edu/group/malaria/start.html>.

Cromosome(s)[a]	Size (Mb)	Laboratory	Funding[b]	Status (as of 7/99)[c]
1	0.8	Sanger Centre	Wellcome Trust	Closure
2	0.95	TIGR/NMRC	NIAID, DoD	Completed (Gardner *et al.*, 1998)
3	1.1	Sanger Centre	Wellcome Trust	Completed (Bowman *et al.*, in press)
4	1.4	Sanger Centre	Wellcome Trust	Closure
5-8	1.6	Sanger Centre	Wellcome Trust	Sequencing
9	1.8	Sanger Centre	Wellcome Trust	Sequencing
10	2.1	TIGR/NMRC	NIAID, DoD	Sequencing
11	2.3	TIGR/NMRC	NIAID, DoD	Closure
12	2.5	Stanford University	BWF	Closure
13	3.2	Sanger Centre	Wellcome Trust	Sequencing
14	3.4	TIGR/NMRC	BWF, DoD	Closure

Sequencing of the first *P. falciparum* chromosome

At the beginning of the Malaria Genome Sequencing Project, *P. falciparum* clone 3D7 was chosen for sequencing because it can complete all stages of the life cycle, was used in a genetic cross (Walliker *et al.*, 1987), and had been used in the Wellcome Trust Malaria Genome Mapping Project (Foster and Thompson, 1995). The TIGR/NMRC group began a pilot project to sequence chromosome 2, which was selected because it could be easily resolved on pulsed-field gels, and being about 1 Mb in size it was not too large to present unsurmountable difficulties in assembly or gap closure. *P. falciparum* chromosomes were resolved on preparative pulsed-field gels and the chromosome 2 bands from several gels were cut out, adjusted to 0.3 M sodium acetate to prevent melting of the AT-rich DNA, and digested with agarose. The DNA was sheared by nebulization and a shotgun library was prepared in pUC18 as described (Fleischmann *et al.*, 1995) except that treatment with *E. coli* DNA polymerase I was performed after the second ligation step to close nicks prior to electroporation. During all steps of the library construction process, the exposure of the DNA to UV light was minimized to avoid damage to the DNA that would reduce the cloning efficiency, particularly of the very AT-rich intergenic sequences. In addition, to prevent generation of non-randomness, the library was not amplified prior to sequencing. Rather, the ligation mixtures were stored at –20°C, and as needed aliquots were electroporated into DH10B cells and spread on ampicillin diffusion plates. The shotgun library contained 1×10^5 recombinants and had an average insert size of 1.6 kb.

Initial sequencing was done with dye-primer chemistry used previously to sequence *H. influenzae* and the other microbial genomes. However, when sequencing the *P. falciparum* clones we observed an apparent artifact with the dye-primer chemistry that resulted in runs of G nucleotide base calls to be incorrectly made following long runs of AT-rich sequence. The artifact did not occur when FS+ dye-terminator chemistry was used on the same template DNAs, and the dye-terminator chemistry also produced significantly longer sequence reads than the dye-primer chemistry. Therefore the rest of the random-phase sequencing was performed using the dye-terminator chemistry. Over 23,000 individual sequences were collected, which was equivalent to about 10x coverage of the chromosome. This is greater coverage than is normally done in a shotgun project, but the excess coverage was thought to be necessary to compensate for the presence of non-chromosome 2 DNA in the library arising from the pulsed-field gel purification of the DNA, and for the expected non-randomness of the shotgun library due to the AT-rich inserts.

The sequence reads were assembled using a version of TIGR Assembler (Sutton *et al.*, 1995) that was extensively modified to assemble the AT-rich and repeat-rich *Plasmodium* sequences. TIGR Assembler identifies and aligns overlapping fragments in two steps. The initial step in assembly is to locate all *n*-mer oligonucleotides shared between fragment pairs. The software views all fragment pairs with a high degree of *n*-mer similarity as potentially overlapping, and in the second step the Smith-Waterman method is used to align the fragments. In the bacterial genome projects the value of *n* used was typically 10-12 nucleotides. However, using n=10 with AT-rich *Plasmodium* DNA resulted in incorrect identification of thousands of potential fragment overlaps, so that the program spent an inordinate amount of time attempting to align the spurious matches. Increasing *n* from 10 to 32 much reduced this problem and significantly lowered the time required for assembly.

After the assembly, 610 contigs were obtained and the largest contigwas 50 kb. Neighboring contigs were identified and ordered by the program GROUPER, which searches for plasmid templates with forward and reverse reads in different contigs (clone links), and for overlapping contigs that failed to merge under the stringent overlap criteria required by TIGR Assembler (grasta links). Contigs within a group are separated by sequence gaps which can be closed by primer walking on the templates identified as clone links, or by editing of the termini of contigs with grasta links. The ends of groups represent physical gaps for which no shotgun clone could be identified. Ten groups of 114 contigs were localized on the chromosome by comparison to STS markers (Lanzer *et al.*, 1993). Closure of physical and sequence gaps used approaches described previously (Fleischmann *et al.*, 1995), with a few modifications to compensate for the AT-richness of the DNA. To close the 9 physical gaps in the central region of the chromosome, PCR reactions using genomic DNA as template were performed with primers from the ends of adjacent groups. PCR products were purified and sequenced using dye-terminator chemistry. This process closed 3 physical gaps immediately, but PCR products from 2 gaps contained very AT-rich sequence which could not be sequenced completely, and remained as sequence gaps. Those physical gaps for which PCR products could not be obtained in the first step were reasoned to be too large for PCR, and to contain one or more of the unlocalized groups. We therefore performed combinatorial PCRs with one primer from the end of a localized group and the second primer from the ends of all free groups larger than 2.5 kb. Two gaps were closed by the combinatoral strategy. Finally, 1 physical gap was closed after editing and reassembly, and another gap was closed by sequencing of a 'missing mate' (i.e., resequencing of a clone for which either the forward or reverse sequencing reaction had failed during the random phase). Five methods were used to close sequence gaps. For contigs which overlapped but had not been merged during assembly, editing and resequencing were per-

formed to close the gaps. Many sequence gaps were caused by artifacts in dye-primer reactions, particularly in extremely AT-rich areas. Long homopolymer stretches of up to 50 consecutive A or T residues also caused the sequence quality to decline downstream of the homopolymer region. These artifacts either prevented the merging of overlapping contigs or produced short sequences that did not extend to the neighboring contig. Some of these problem areas could be solved by trimming of the low quality sequence that prevented merging of the contigs. For other gaps, templates from short or low-quality dye-primer reactions in the vicinity of sequence gaps were identified and resequenced with dye-terminator chemistry; the longer reads of high-quality sequence provided by the dye-terminator reactions was sufficient to close many gaps. For those gaps that remained, primer walking on plasmid templates linking adjacent contigs was used. Finally, there were 5 sequence gaps that could not be closed by the above methods because the sequence was too AT-rich for primer synthesis and walking. To close these gaps, the artificial transposon AT-2 (Devine and Boeke, 1994) was inserted into one of the templates spanning each sequence gap, multiple subclones of each template were sequenced using transposon-specific primers, and the sequences were assembled to close the gap. The chromosome 2 sequence was edited manually using TIGR Editor, and where necessary additional sequencing reactions were performed to improve coverage and resolve sequence ambiguities. One major concern, given the well-known propensity for AT-rich *P. falciparum* sequences to rearrange in *E. coli*, was whether the assembled sequence was an accurate representation of the genomic sequence. To independently confirm the colinearity of the assembled sequence and genomic DNA, *Nhe*I and *Bam*HI optical restriction maps of chromosome 2 DNA were prepared and compared with restriction maps predicted from the sequence (Jing *et al.*, 1999). The relative error of predicted and observed fragment sizes was less than 6%, which proved that there were no major rearrangements in the assembled sequence.

Annotation of *P. falciparum* chromosome 2

Annotation of the chromosome 2 sequence followed the procedures used previously during the annotation of other genomes, including BLAST searching of all open reading frames (ORFs) against a protein sequence database. In addition, to assist in defining the intron/exon boundaries, a new eukaryotic gene finding program was developed specifically for use in this project (Salzberg *et al.*, 1999). This program, GlimmerM, was trained on a set of 117 *P. falciparum* sequences taken from Genbank. Gene models based on the GlimmerM predictions, the similarity of the ORFs to known proteins, and prediction of putative signal peptides and transmembrane domains were constructed.

Chromosome 2 of *P. falciparum* is 947,103 bp in length and 80.2% A+T (Gardner *et al.*, 1998). It possesses typical eukaryotic telomeres and subtelomeric regions containing several kb of rep20 tandem repeats, variant antigen genes (*var*), and a potential new family of variant surface antigens related to the RIF-1 elements (repetitive interspersed family) (Weber, 1988). The large central region encodes many single copy genes and several genes that are tandemly repeated (Fig. 1). Two hundred and nine protein-encoding genes and a gene encoding $tRNA^{Glu}$ were predicted on chromosome 2, giving a gene density of one gene per 4.5 kb, which is significantly lower than in yeast (one gene per 2 kb) but higher than in *C. elegans* (one gene per 7 kb). It was estimated that 43 of the 209 protein-encoding genes contained at least one intron, with most such genes consisting of 2 or 3 exons. Two genes, however, contained 8 exons. Extrapolation of the chromosome 2 data to the entire 28 Mb *P. falciparum* genome suggests that it contains 6,200 genes, 2,600 of which may contain introns. Thus, in terms of intron content and gene density the *P. falciparum* genome appears to be intermediate between the compact yeast genome and the intron-rich genomes of multicellular eukaryotes.

Of the 209 protein encoding genes, only 87 (42%) appeared to have homologs outside *Plasmodium*, suggesting that almost 60% of the genes encoded on this chromosome are so far 'unique' to *Plasmodium*. The proportion of unique genes is almost 2-fold greater than has been observed in other organisms, and confirms that there is much biology to be uncovered in future studies of this parasite. As sequencing of other related parasites proceeds, some of these proteins will undoubtedly be found to have homologs in apicomplexans such as *Toxoplasma* (Ajioka *et al.*, 1998) and *Eimeria*, and hence may be found to be characteristic of apicomplexan parasites. Most of the remaining unidentified proteins on chromosome 2 were predicted to consist primarily of non-globular domains, i.e. domains that are composed of low complexity sequences that do not form compact folded structures (Wootton and Federhen, 1996). The abundance of non-globular domains or proteins in *Plasmodium* was very unusual, and was about half that observed in *S. cerevisiae*, *C. elegans*, and humans. In addition, 13 proteins contained large regions (>30 amino acids) with predicted non-globular structure inserted directly into globular domains, a phenomenon so far unique to *Plasmodium*. These non-globular insertions did not exhibit the AT-bias typical of introns, were not flanked by consensus splice sites, and based on RT-PCR analysis of several genes encoding non-globular domains, were likely to be expressed in the proteins. The abundance of the non-globular domains in *Plasmodium* proteins suggests that they provide as yet unknown selective advantages to the parasite. Study of these proteins containing non-globular inserts may also provide new insight into the general principles of protein folding.

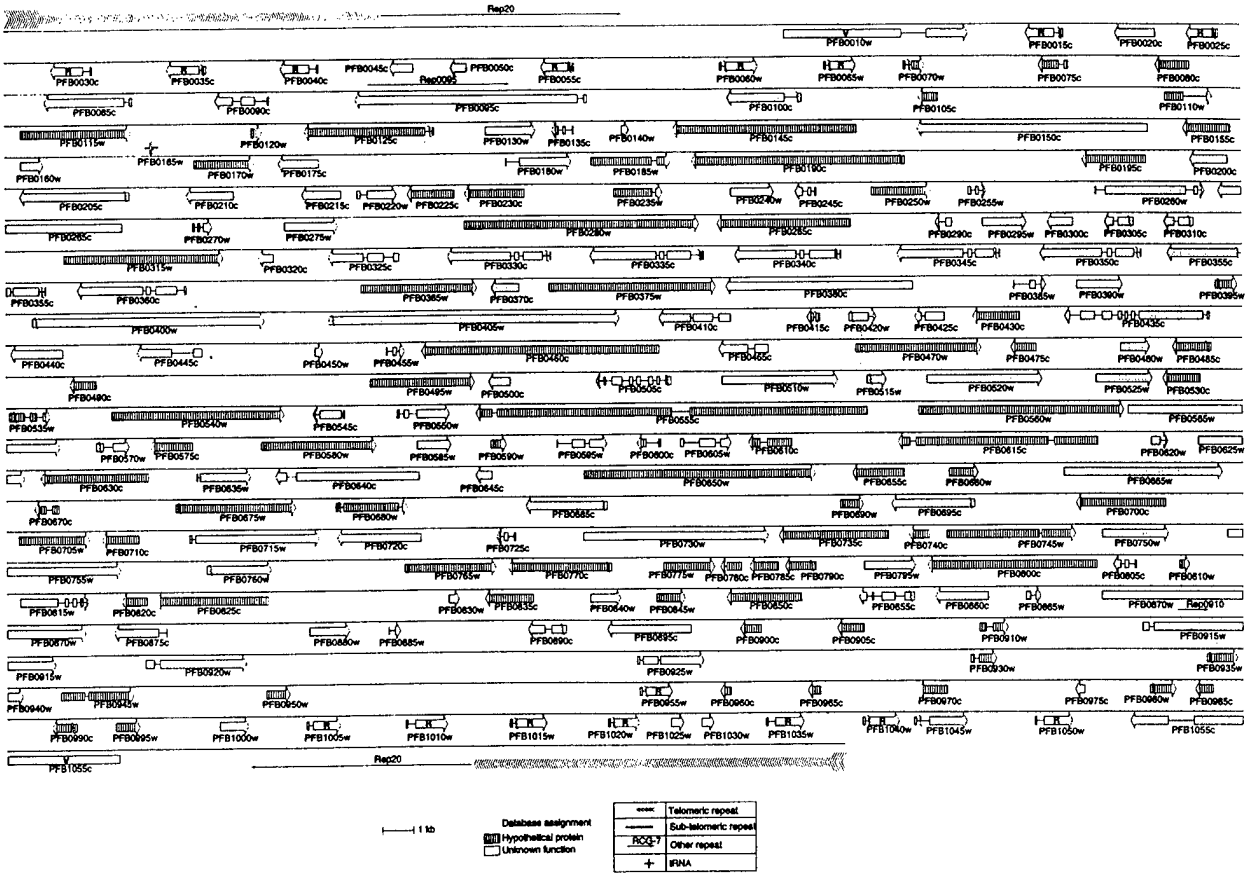

Fig. 1. Map of *P. falciparum* chromosome 2 (clone 3D7). Exons are shown as boxes or arrows, with introns represented by thin lines connecting the exons. Other features such as telomeric and subtelomeric repeats are indicated as shown in the legend. Chromosome 2 genes with similarity to known genes in the sequence databases and for which putative functional assignments could be made are stippled; hypothetical genes with no detectable similarity to known genes are indicated by vertical stripes; genes with similarity to previously sequenced genes of unknown function are indicated as open arrows. The rifin and *var* genes are labeled with 'R' and 'V', respectively. Genes were given systematic names using a scheme similar to that devised for the *S. cereviseae* genome (Mewes *et al.*, 1997). For a complete description of the genes encoded on chromosome 2, including details of functional assignments, see Gardner *et al.* (1998).

Most of the 87 evolutionarily-conserved proteins encoded on chromosome 2 show the greatest similarity to eukaryotic homologs or belong to specifically eukaryotic protein families. Many of these genes code for proteins that participate in replication, repair, transcription, or translation, and include the origin recognition complex subunit 5, two proteins involved in excision repair proteins, several proteins involved in chromatin dynamics, RNA-binding proteins, and a putative transcription factor. Other evolutionarily conserved proteins are involved in secretion, such as the SEC61 gamma subunit, the coated pit coatamer subunit, and syntaxin, suggesting early emergence of the eukaryotic secretory system. Five proteins contained DnaJ domains; in other organisms DnaJ proteins have been shown to act as cofactors for the HSP70-type molecular chaperones and to participate in a variety of processes such as protein folding and trafficking, complex assembly, organelle biogenesis, and initiation of translation (Cyr *et al.*, 1994). Chromosome 2 encodes 90 predicted membrane proteins, some of which appear to be transporters of amino acids or sugars. Five putative protein kinases were also identified, suggesting that the *P. falciparum* genome may encode about 150 protein kinases. This prominence of regulators is in striking contrast to the situation in bacterial pathogens, which appear to have shed most of the regulatory systems, and is probably a reflection of the complex life cycle. For example, phosphorylation and dephosphorylation reactions are known to be involved in the development and sexual differentiation of malaria parasites (Bracchi *et al.*, 1996). A cluster of 8 tandemly arranged genes encoding putative proteases was also found; 3 of these genes were known previously and were called SERAs (SErine Repeat Antigens). The expansion of this protease gene family suggests an important function, possibly in merozoite release from schizonts or processing of merozoite surface proteins.

While most of the evolutionarily conserved proteins were more similar to eukaryotic homologs, 16 proteins were significantly more similar to bacterial homologs and 4 other proteins were the first eukaryotic representatives of conserved bacterial protein families. These proteins may have been transferred to the nuclear genome from an organellar genome after the divergence of the apicomplexa from the other eukaryotic lineages. Several of these proteins contained N-terminal sequences that resembled organellar import peptides, which suggested that these proteins may be imported into and function within either the apicoplast or the mitochondrion. Of particular interest were 3 genes encoding proteins involved in fatty acid metabolism. One of these proteins, 3-ketoacyl-ACP synthase III (FabH), catalyzes the condensation of acetyl-CoA and malonyl-ACP in Type II (dissociated) fatty acid synthase systems. Type II synthase systems are restricted to bacteria and the plastids of plants, and the discovery of a Type II fatty acid synthase system in *Plasmodium* reinforced previous hypotheses that the apicoplast contains plant-like metabolic pathways distinct from those of the host (Wilson *et al.*, 1991; Slabas and Fawcett, 1992). Some of the biochemical processes that occur within this organelle may therefore be good drug targets (Soldati, 1999). Recent work has confirmed that at least some of the predicted import peptides can direct translocation of reporter proteins into the apicoplast in *Toxoplasma*, and in addition, that thiolactomycin, a specific inhibitor of bacterial FabH, can inhibit the growth of *P. falciparum* in vitro (Waller *et al.*, 1998).

As mentioned previously, more than half of all proteins encoded on chromosome 2 did not have detectable homologs in other species. Many of the *Plasmodium* specific genes were located in the subtelomeric regions of the chromosome. Two members of the *var* gene family were identified on chromosome 2, one in each sub-telomeric region. The *var* genes encode large proteins, collectively known as PfEMP1s, that are located on the surface of infected red cells, exhibit extensive sequence diversity, and are involved in antigenic variation, cytoadherence, and rosetting (Baruch *et al.*, 1995; Smith *et al.*, 1995; Su *et al.*, 1995; Rowe *et al.*, 1997). Most *var* genes are located in sub-telomeric regions, and *var* gene diversity is thought to be generated by recombination between alleles, a process which might be facilitated by the sub-telomeric repeats (Rubio *et al.*, 1996). Six small ORFs that had similarity to *var* sequences were also found in the sub-telomeric regions. Five of these ORFs resembled the *var* exon II cDNAs or the Pf60.1 sequences that were reported previously (Su *et al.*, 1995; Bonnefoy *et al.*, 1997). However, the largest gene family identified on chromosome 2 encoded proteins of 27-35 kD that were named rifins, after the RIF-1 repetitive elements (Weber, 1988). These proteins contained a N-terminal signal sequence, a central region of variable length and an amino acid sequence containing conserved cysteine residues, a transmembrane domain, and a C-terminus rich in basic amino acids, and were predicted to be expressed on the surface of infected red cells. All eighteen of the rifin genes were in the subtelomeric regions, centromere proximal to the *var* genes. Clusters of rifin genes have been detected on other chromosomes (Cheng *et al.*, 1998), and if the number of rifins found on chromosome 2 is representative of the other chromosomes, the *P. falciparum* genome may contain more than 500 rifin genes. While the function of the rifins is not known, the extensive sequence diversity of the rifins suggests that, like the *var* gene products, they may be clonally variant. Further studies are underway in a number of laboratories to confirm the subcellular localization of the rifins and to determine their function.

Future prospects

The completion of the first *P. falciparum* chromosome and the rapid progress being made by all three genome centers on the remaining chromosomes (Table 1) suggests that the entire *P. falciparum* genome will be completed within 2-3 years. In fact, it is quite likely that most of the parasite's genes will have been identified within 18-24 months, with the additional time being spent on the closing of gaps in the sequence. Ideally, the completion of the *P. falciparum* genome sequence will be followed by the sequencing of a second *Plasmodium* species so as to provide valuable comparative information. The human parasite *P. vivax* and several rodent malaria parasites used as model systems for vaccine and drug development are currently viewed as candidates for sequencing. In addition, information derived from expressed sequence tag (EST) or genome sequencing projects for other apicomplexa such as *Toxoplasma* (Ajioka *et al.*, 1998) will help to identify parasite-specific metabolic pathways that will be useful for development of new drugs against these organisms. Recent technological advances such as the stable transfection of several *Plasmodium* species (van Dijk *et al.*, 1995; Wu *et al.*, 1995; Crabb and Cowman, 1996; van der Wel *et al.*, 1997) and the ability to knock-out specific genes (Menard *et al.*, 1996; Crabb *et al.*, 1997), and the development of microarray technologies for global measurements of gene expression (Schena *et al.*, 1995), will help in the interpretation of the genome sequence. This is important in view of the fact that less than one-half of all the genes identified on the first *P. falciparum* chromosome to be sequenced could be assigned functional roles. Clearly, there is much exciting research to be done and researchers studying *Plasmodium* and related parasites can look forward to Bloom's post-genomic era of microbe biology.

Acknowledgements

The Malaria Genome Sequencing Project is supported by The Wellcome Trust, the US Department of Defense, The Burroughs Wellcome Fund and the National Institutes of Health. This work

was supported by a supplement to NIH grant R01-AI40125-01; the Naval Medical Research and Development Command work units 61102A.S13.00101.BFX.1431, 612787A.870.00101.EFX.1432, 623002A.810.00101.HFX.1433 and STEP C611102A0101 BCX; Department of the Army Cooperative Agreement DAMD17-98-2-8005; and NIH-NMRC interagency agreement No. Y1AI-6091-01. The opinions and assertions herein are those of the authors and are not to be construed as official or as reflecting the views of the US Navy or Department of the Army.

References

Ajioka JW, Boothroyd JC, *et al.* (1998). Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa. Genome Res 8: 18-28.

Baruch DI, Pasloske BL, *et al.* (1995). Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 82: 77-87.

Bloom BR (1995). A microbial minimalist. Nature 378: 236.

Bonnefoy S, Bischoff E, *et al.* (1997). Evidence for distinct prototype sequences within the *Plasmodium falciparum* Pf60 multigene family. Mol Biochem Parasitol 87: 1-11.

Bowman S, Lawson D, *et al.* (1999). The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature (in press).

Bracchi V, Langsley G, *et al.* (1996). PfKIN, an SNF1 type protein kinase of *Plasmodium falciparum* predominantly expressed in gametocytes. Mol Biochem Parasitol 76: 299-303.

Blattner FR, Plunket G, *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1474.

Cheng Q, Cloonan N, *et al.* (1998). *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. Mol Biochem Parasitol 97: 161-176.

Cole ST, Brosch R, *et al.* (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537-544.

Crabb BS, Cooke BM, *et al.* (1997). Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. Cell 89: 287-296.

Crabb BS, Cowman AF (1996). Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*. Proc Natl Acad Sci USA 93: 7289-7294.

Cyr DM, Langer T, *et al.* (1994). DnaJ-like proteins: molecular chaperones and specific regulators of Hsp70. Trends Biochem Sci 19: 176-181.

Dame JB, Arnot DE, *et al.* (1996). Current status of the *Plasmodium falciparum* genome project. Mol Biochem Parasitol 79: 1-12.

Devine SE, Boeke JD (1994). Efficient integration of artifical transposons into plasmid targets in vitro: a useful tool for DNA mapping, sequencing, and genetic analysis. Nucleic Acids Res 22: 3765-3772.

Fleischmann RD, Adams MD, *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496-512.

Foster J, Thompson J (1995). The *Plasmodium falciparum* genome project: a resource for researchers. Parasitol Today 11: 1-4.

Fraser CM, Casjens S, *et al.* (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390: 580-586.

Fraser CM, Norris SJ, *et al.* (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281: 375-388.

Gardner MJ, Tettelin H, *et al.* (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science 282: 1126-1132.

Hoffman SL, Bancroft WH, *et al.* (1997). Funding for malaria genome sequencing. Nature 387: 647.

Jing J, Aston C, *et al.* (1999). Optical mapping of *Plasmodium falciparum* chromosome 2. Genome Res 9: 175-181.

Lanzer M, de Bruin D, *et al.* (1993). Transcriptional differences in polymorphic and conserved domains of a completed cloned *P. falciparum* chromosome. Nature 361: 654-657.

Menard R, *et al.* (1996). Circumsporozoite protein is required for development of malaria sporozoites in mosquitos. Nature 385: 336-340.

Mewes HW, Albermann K, *et al.* (1997). Overview of the yeast genome. Nature 387: 7S-65S.

Rowe JA, Moulds JM, *et al.* (1997). *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. Nature 388: 292-295.

Rubio JP, Thompson JK, *et al.* (1996). The *var* genes of *Plasmodium falciparum* are located in the subtelomeric region of most chromosomes. EMBO J 15: 4069-4077.

Salzberg SL, Pertea M, *et al.* (1999). Interpolated Markov models for eukaryotic gene finding. Genomics 59: 24-31.

Schena M, Shalon D, *et al.* (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470.

Slabas AR, Fawcett T (1992). The biochemistry and molecular biology of plant lipid biosynthesis. Plant Mol Biol 19: 169-191.

Smith JD, Chitnis CE, *et al.* (1995). Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell 82: 101-110.

Soldati D (1999). The apicoplast as a potential therapeutic target in *Toxoplasma* and other apicomplexan parasites. Parasitol Today 15: 5-7.

Stephens RS, Kalman S, *et al.* (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282: 754-759.

Su Z, Heatwole VM, *et al.* (1995). The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. Cell 82: 89-100.

Sutton GS, White O, *et al.* (1995). TIGR Assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1: 9-19.

Tomb JF, White O, *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388: 539-547.

van der Wel AM, Tomas AM, *et al.* (1997). Transfection of the primate malaria parasite *Plasmodium knowlesi* using entirely heterologous constructs. J Exp Med 185: 1499-1503.

van Dijk MR, Waters AP, *et al.* (1995). Stable transfection of malaria parasite blood stages. Science 268: 1358-1362.

Waller RF, Keeling PJ, *et al.* (1998). Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. Proc Natl Acad Sci USA 95: 12352-12357.

Walliker D, Quayki I, *et al.* (1987). Genetic analysis of the human malaria parasite *Plasmodium falciparum*. Science 236: 1661-1666.

Weber JL (1988). Interspersed repetitive DNA from *Plasmodium falciparum*. Mol Biochem Parasitol 29: 117-124.

Wilson RJM, Gardner MJ, *et al.* (1991). Have malaria parasites three genomes? Parasitol Today 7: 134-136.

Wootton JC and Federhen S (1996). Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266: 554-571.

Wu Y, Sifri CD, *et al.* (1995). Transfection of *Plasmodium falciparum* within human red cells. Proc Natl Acad Sci USA 92: 973-977.

S. cerevisiae and the latter, identified genes unique to the latter that may prove to be restricted to metazoans and play a role in multicellularity.

14. Bruccoleri RE, Dougherty TJ, Davison DB: **Concordance analysis of microbial genomes.** *Nucleic Acids Res* 1998, **26**:4482-4486.
• A rare example showing the importance of combining *in silico* and experimental approaches to hypothesis testing. Using the 'species-filter' approach in combination with experimental approaches the authors search for drug targets – essential genes restricted to pathogenic prokaryotes. Useful software for comparing shared and unshared genes between complete genomes available on the internet is described.

15. Allsop AE: **New antibiotic discovery, novel screens, novel targets and impact of microbial genomics.** *Curr Opin Microbiol* 1998, 1:530-534.

16. Allsop AE: **Bacterial genome sequencing and drug discovery.** *Curr Opin Biotechnol* 1998, **9**:637-642.

17. Hood DW, Deadman ME, Allen T, Masoud H, Martin A, Brisson JR, Fleischmann R, Venter JC, Richards JC, Moxon ER: **Use of the complete genome sequence information of *Haemophilus influenzae* strain Rd to investigate lipopolysaccharide biosynthesis.** *Mol Microbiol* 1996, **22**:951-965.

18. Kalman S, Mitchell W, Marathe R, Lammel C, Fan LL, Hyman RW, Olinger L, Grimwood L, Davis RW, Stephens RS: **Comparative genomes of *Chlamydia pneumoniae* and *C-trachomatis*.** *Nat Genet* 1999, **21**:385-389.

19. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG: **The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria.** *Nature* 1998, **396**:133-140.

20. Bishai W: **The *Mycobacterium* tuberculosis genomic sequence: anatomy of a master adaptor.** *Trends Microbiol* 1998, **6**:464-465.

21. Gelbart WM: **Databases in genomic research.** *Science* 1998, **282**:659-661.

22. **Kyoto Encyclopedia of Genes and Genomes (KEGG).** http://star.scl.kyoto-u.ac.jp/kegg/

23. **What is there (WIT)? Interactive metabolic reconstruction on the web.** http://wit.mcs.anl.gov/WIT2/

24. Normile D: **Building working cells 'in silico'.** *Science* 1999, **284**:80-81.
•• The concept of '*in silico*' cells is at present an underdeveloped yet exciting possibility that holds the promise of allowing researchers to combine all aspects of cell biology into a single system that 'lives' within a computer. From genes, to metabolic pathways, to information on stochiometric ratios of protein products within a cell – once sufficient information has been collected and entered into these model cells, this will be a technology to watch in the future (see [25]).

25. Schaff J, Fink CC, Slepchenko B, Carson JH, Loew LM: **A general computational framework for modeling cellular structure and function.** *Biophys J* 1997, **73**:1135-1146.

26. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21**:33-37.
•• Review of the subject in a special issue devoted to advances in microarray technology.

27. Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, Davis RW: **Genomic profiling of drug sensitivities via induced haploinsufficiency.** *Nat Genet* 1999, **21**:278-283.

28. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ, Davis RW: **Direct allelic variation scanning of the yeast genome [see comments].** *Science* 1998, **281**:1194-1197.

29. Spratt BG, Maiden MC: **Bacterial population genetics, evolution and epidemiology.** *Philos Trans R Soc Lond Biol* 1999, **354**:701-710.

30. Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM: **Comparative genomics of BCG vaccines by whole-genome DNA microarray.** *Science* 1999, **284**:1520-1523.
•• A premier example of how microarray technology can be used to survey for heritable genetic variation that may provide the genetic basis for important phenotypes. In this case, genetic variation in tuberculosis vaccine strains that may explain observed temporal and geographic differences in vaccine effectiveness.

31. Sokurenko EV, Hasty DL, Dykhuizen DE: **Pathoadaptive mutations: gene loss and variation in bacterial pathogens.** *Trends Microbiol* 1999, **7**:191-195.

32. Groisman EA, Ochman H: **How to become a pathogen.** *Trends Microbiol* 1994, **2**:289-294.

33. Gordon SV, Heym B, Parkhill J, Barrell B, Cole ST: **New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv.** *Microbiology* 1999, **145**:881-892.

34. Isaksson A, Landegren U: **Accessing genomic information: alternatives to PCR.** *Curr Opin Biotechnol* 1999, **10**:11-15.

35. Pennisi E: **Is it time to uproot the tree of life?** *Science* 1999, **284**:1305-1307.
• See papers [36–41]. Excellent synopsis of the current debate over how much horizontal gene transfer contributes to the evolution of bacterial genomes.

36. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nat Genet* 1999, **21**:108-110.

37. Teichmann SA, Mitchison G: **Is there a phylogenetic signal in prokaryote proteins?** *J Mol Evol* 1999, **49**:98-107.

38. Lake JA, Jain R, Rivera MC: **Mix and match in the tree of life.** *Science* 1999, **283**:2027-2028.

39. Baumler AJ: **The record of horizontal gene transfer in *Salmonella*.** *Trends Microbiol* 1997, **5**:318-322.

40. Conner CP, Heithoff DM, Julio SM, Sinsheimer RL, Mahan MJ: **Differential patterns of acquired virulence genes distinguish Salmonella strains.** *Proc Natl Acad Sci USA* 1998, **95**:4641-4645.

41. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci USA* 1999, **96**:3801-3806.

42. Groisman EA, Ochman H: **How Salmonella became a pathogen.** *Trends Microbiol* 1997, **5**:343-349.

43. Malakoff D: **NIH urged to fund centers to merge computing and biology.** *Science* 1999, **284**:1742.
•• Evidence that future funding will place emphasis on the establishment of 'centers of excellence' dedicated to projects that merge computational and empirical approaches. As so many genome projects now focus on pathogenic species, this initiative will hopefully have an important impact on the study of pathogenesis.

44. Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C *et al.*: **Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*.** *Science* 1998, **282**:1126-1132.

45. Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S *et al.*: **Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes.** *Proc Natl Acad Sci USA* 1999, **96**:2902-2906.

46. **The Sanger Center.** http://www.sanger.ac.uk

47. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al.*: **The minimal gene complement of *Mycoplasma genitalium*.** *Science* 1995, **270**:397-403.

48. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*.** *Nucleic Acids Res* 1996, **24**:4420-4449.

49. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA *et al.*: **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, **388**:539-547.

50. Cole ST: **Comparative mycobacterial genomics.** *Curr Opin Microbiol* 1998, 1:567-571.

51. **The Institute for Genome Research (TIGR).** http://www.tigr.org

52. **Advanced Center for Genome Technology (ACGT).** http://dna1.chem.ou.edu/

The genome of the malaria parasite

Malcolm J Gardner

The genome of the human malaria parasite *Plasmodium falciparum* is being sequenced by an international consortium. Two of the parasite's 14 chromosomes have been completed and several other chromosomes are nearly finished. Even at this early stage of the project, analysis of the genome sequence has provided promising new leads for drug and vaccine development.

Addresses
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA; e-mail: gardner@tigr.org

Current Opinion in Genetics & Development 1999, **9**:704–708

0959-437X/99/$ – see front matter

Abbreviations

CTL	cytotoxic T lymphocyte
EST	expressed sequence tag
GST	gene sequence tag
STS	sequence-tagged site
YAC	yeast artificial chromosome

Introduction

Over one-third of the world's population is at risk of contracting malaria, a mosquito-borne disease caused by apicomplexan parasites of the genus *Plasmodium*. There are ~300–500 million new cases and ~1.5–2.7 million deaths from malaria annually. Most deaths due to malaria occur among children in sub-Saharan Africa [1]. At present there is no effective, practical vaccine that can be used to prevent malaria, and although there are effective antimalarial drugs, resistance to one of more of these drugs has developed in many parts of the world. Development of new drugs and vaccines has been only moderately successful, limited by the financial resources that are available and the difficulty of working with a complex intracellular parasite. (A comprehensive collection of review articles on all aspects of *Plasmodium* biology can be found here [2•].)

Completion of the first microbial genome sequences demonstrated the benefits that accrue from genome sequencing [3]. For a pathogenic organism, the genome sequence provides the sequence of every potential drug or vaccine target; for difficult to study organisms like *Plasmodium*, sequencing of the genome may be the only way to identify these targets. The *Plasmodium falciparum* genome is approximately 28 megabase pairs (Mb) in length and contains 14 chromosomes ranging in size from ~0.6–3.4 Mb. Chromosome sizes can vary markedly between wild isolates as a result of recombination events involving the repeat-rich subtelomeric regions of the chromosome. The genome is extremely A+T rich (~80%), which might account for the instability of large fragments of *P. falciparum* DNA in *E. coli*. The DNA is more stable in yeast; large insert yeast artificial chromosome (YAC) libraries have been constructed and used to generate STS (sequence-tagged site) maps of most of the chromosomes [4]. In addition, a linkage map of the genome consisting of more than 900 microsatellite markers and having a resolution of 30 kb has been produced [5••]. Expressed sequence tags (ESTs) from blood stage parasites and gene sequence tags (GSTs) have also been prepared [6,7]. Techniques for manipulation of the genome have been developed including stable transfection and gene knockouts [8••]. This review summarizes recent progress in the sequencing of the *P. falciparum* genome, and outlines how the genome sequence information produced in this effort is contributing to the development of new drugs and vaccines against malaria.

The *Plasmodium falciparum* genome sequencing project

P. falciparum is the most lethal of the four *Plasmodium* species that cause malaria in humans. Fortunately, all stages of the *P. falciparum* life cycle can be maintained in the laboratory, blood stages can be cultured routinely, and cloned parasites are available. In late 1996, a consortium of funding agencies, genome centers, and malaria investigators was formed to sequence the *Plasmodium falciparum* genome [9,10]. A strategy was adopted whereby individual chromosomes assigned to each genome center were resolved by pulsed field gel electrophoresis and subjected to shotgun sequencing. STS markers [4], the microsatellite linkage map [5••,11], and optical restriction maps [12••,13] of the chromosomes were used for ordering of the contiguous sequences during the gap closure phase and for verification of the final sequence assembly. Chromosomes 2 and 3, which comprise about 7% of the genome, have been completed [14••,15••]; preliminary data at various stages of completion are available for the remaining chromosomes (Table 1). One difficulty faced by the sequencing groups was the identification of genes in the A+T-rich sequence. Gene finding algorithms developed for higher eukaryotes, which have a much lower gene density than *Plasmodium*, were not optimal for the prediction of coding regions in *Plasmodium* DNA, and prokaryotic gene finders were unable to predict introns. GlimmerM gene finding software was developed during the chromosome 2 project; it uses interpolated Markov models constructed from a training set of well-characterized genes for prediction of coding regions and a separate module for prediction of splice sites [16].

The chromosome sequences revealed that 20–30 kb of each chromosome end was composed of telomeric, rep20, and other repeats [14••,15••]. Centromeric to these repeats, members of multigene families involved in antigenic variation and or pathogenesis were found [17•], including *var* genes that encode the PfEMP1 proteins [18–21], open reading frames with similarity to the 3′ exon of *var* genes

Table 1

Web sites related to the malaria genome sequencing project.

Web site	Content	URL
P. falciparum chromosomes 2, 10, 11, 14, TIGR/Naval Medical Research Center	Chromosome 2 annotation [14••] Preliminary data	http://www.tigr.org/tdb/mdb/pfdb/pfdb.html
P. falciparum chromosomes 1, 3, 4, 5–9, 13, The Sanger Centre	Chromosome 3 annotation [15••] Preliminary data	http://www.sanger.ac.uk/Projects/P_falciparum/
P. falciparum chromosome 12, Stanford University	Preliminary data for chromosome 12	http://sequence www.stanford.edu/group/malaria/index.html
P. falciparum Gene Sequence, Tag Project University of Florida	A collection of ESTs and GSTs for *P. falciparum* [6,7]	http://parasite.arf.ufl.edu/malaria.html
Malaria Database, Monash University, Walter and Eliza Hall Institute	A collection of genetic information on malaria parasites	http://www.wehi.edu.au/MalDB-www/who.html
Malaria Genetics and Genomics, National Center for Biotechnology Information (NCBI)	BLAST searches on Apicomplexan sequence data, including *P. falciparum*; *P. falciparum* linkage maps, etc.	http://www.ncbi.nlm.nih.gov/Malaria/
Parasite Genomes Blast Server, European Bioinformatics Institute	BLAST searches on sequence data from many parasites, including *Plasmodium*	http://www.embl-ebi.ac.uk/parasites/parasite_blast_server.html
Malaria Foundation	General information on malaria and many links to malaria-related sites	http://www.malaria.org/index.htm
TIGR Microbial Database	A comprehensive listing of microbial genome projects	http://www.tigr.org/tdb/mdb/mdb.html

BLAST, basic local alignment search tool; dbEST, database of expressed sequence tags; GSTs, genome sequence tags.

that may represent a distinct gene family [22], and members of the *rif* and *STEVOR* gene families (see below). Gene density was about 1 gene per 4.7 kb and almost one-half of genes were predicted to contain introns. Depending upon the methods used for annotation, up to two-thirds of the genes identified had no detectable orthologs in other organisms, which suggests that our current understanding of malaria parasite biology is woefully incomplete.

The investment in sequencing of the genome has already paid handsome dividends. A large gene family (*rif*) was identified on chromosome 2 [14••] (the *STEVOR* family was proposed to be a family related to, but distinct from, the *rif* family [23••]). The *rif* genes encoded polypeptides of 27–35 kDa (rifins) that were predicted to be located on the red cell surface and which contained a region variable in length and amino acid sequence. The sequence polymorphism of the rifins, their presumed cell surface localization, and the distribution of the *rif* genes in subtelomeric regions near the *var* genes suggested that rifins might be a new class of variant surface antigen. Laboratory studies have now proven that the rifins are expressed on the surface of the infected red cell and that they are clonally-variant but the function of these proteins has not been determined [24••]. Like the PfEMP1 proteins encoded by the *var* gene family, which mediate cytoadherence and rosetting, the rifins might have a role in host–parasite interaction.

Other major findings included the discovery of genes encoding enzymes of the type II fatty acid biosynthetic pathway that were previously found only in plants and bacteria [14••,25••], a cluster of four genes of unknown function that was repeated on one end of both chromosomes 2 and 3, and the identification of putative centromere sequences [15••]. The predicted centromeres (~2–3 kb in length) were located in the most A+T rich region of each chromosome (>97% A+T), which in both cases were under represented in the plasmid shotgun libraries used for sequencing and were the most difficult regions to sequence. Proof that these regions actually are centromeres awaits improvements in transfection technology; however, if these are centromeres they could be useful for the stable maintenance of minichromosomes in transfected parasites.

Identification of new chemotherapeutic targets using the genome sequence

Investigation of a 35 kb extra-chromosomal DNA with features characteristic of plastid DNA by Wilson and

colleagues [26] led to the identification of an organelle in *Plasmodium*, *Toxoplasma*, and related parasites called the apicoplast [27–30]. Early studies revealed that organellar protein synthesis and DNA replication were targets of antibiotics with antimalarial activity (for reviews see [31,32]). Analysis of the complete sequence of the 35 kb DNA provided few clues to the function of the organelle but, like plastids of higher plants, the organelle was hypothesized to contain biochemical pathways essential for cell survival. If such pathways were parasite specific they would make attractive drug targets.

Because most proteins in the plastids of other organisms are encoded by nuclear DNA and imported into plastids, it was clear that the genes encoding the enzymes of these pathways were to be found in the nuclear genome. Analysis of the genome sequence in conjunction with transfection studies in *Toxoplasma* have led to the identification of nuclear-encoded proteins that are imported into the apicoplast and the amino-terminal sequences that direct the transport of these proteins into the organelle [25••]. The *fab*H gene encoding 3-ketoacyl acyl carrier protein synthase III — an enzyme involved in type II fatty acid biosynthesis — was identified on chromosome 2 in *P. falciparum* and shown to contain a putative apicoplast-targeting peptide. The antibiotic thiolactomycin, which inhibits the orthologous bacterial enzyme, was shown to possess growth-inhibitory activity against *P. falciparum in vitro* [25••].

Most recently, genes encoding enzymes of the non-mevalonate pathway of isoprenoid biosynthesis were identified in preliminary data from the chromosome 14 sequencing project and the enzymes were predicted to be localized in the apicoplast [33••,34]). Inhibitors of one enzyme in the pathway (1-deoxy-D-xylulose 5-phosphate reductoisomerase) were found to inhibit the activity of the recombinant enzyme expressed in bacteria and to possess antiparasite activity *in vitro* and *in vivo*. These examples validate the early interest in plastid-localized pathways as drug targets, and demonstrate the rapidity with which potential drug targets can be identified with genome sequence information.

Investigators searching for new drug targets have also found plant-like biochemical pathways in apicomplexan parasites that may not be located in the apicoplast (e.g. the shikimate pathway) using more conventional approaches [35•,36]. Other potential targets not related to the apicoplast have also been identified via gene sequence information [37]. As the sequencing of the genome proceeds it will be possible to construct an increasingly comprehensive view of parasite metabolism (the 'metabolome'), which should permit the identification of many more novel drug targets. Successful exploitation of these novel targets may reduce reliance on current antimalarials to which resistance has developed and permit the development of multi-drug therapies that may slow the development of resistance in the future.

Genome sequence and vaccine development

The *P. falciparum* genome sequence will also provide the amino acid sequences of all potential vaccine antigens. Characterization of the hundreds or thousands of antigens to be identified from the genome sequence and their formulation into effective vaccines will be a formidable task — one made more difficult by the requirement that each vaccine must elicit the appropriate immune response for targeting of the different stages of the parasite life cycle [38,39]. One proposed approach is to clone individual *P. falciparum* genes or long open reading frames into DNA vaccines, generate antisera to the encoded proteins in mice, and use immunofluorescence assays to determine the expression patterns and subcellular localization of the candidate antigens in the parasite [40••]. Antigens expressed only within infected hepatocytes, which are targeted primarily by CD8+ T cell responses, could be screened via computer algorithms to predict cytotoxic T lymphocyte (CTL) epitopes. The CTL epitopes could be combined into a series of multi-epitope DNA vaccine constructs and multicomponent DNA vaccines encoding many full-length liver stage antigens could also be prepared. Blood stage antigens accessible to antibodies could also be formulated into DNA vaccines. Clinical trials to establish immunogenicity and protective efficacy of the vaccines would follow. Pilot projects using genes from the two completed chromosomes could be used to validate this approach prior to its application on a large scale. Other approaches to the use of genome data for vaccine development are also possible, including scaling-up of the current antigen-by-antigen strategy using rodent malaria orthologs to *P. falciparum* antigens, or targeted expression library immunization techniques [41].

Comparative genomics

Four species of *Plasmodium* are currently known to infect humans. *P. falciparum* is by far the most lethal of the four species, but *P. vivax*, *P. malariae*, and *P. ovale* cause significant morbidity. *P. vivax* is the most prevalent of these and is of increasing concern because of the development of chloroquine resistance. Apart from the sequencing of genes encoding potential vaccine antigens and drug targets, comparatively little molecular biology has been done with these parasites, primarily because they are extremely difficult or impossible to culture continuously *in vitro* [42] and must be maintained in primates. Carlton *et al.* [43•] have produced karyotype maps of the three other human *Plasmodia*. Like *P. falciparum*, these species appear to have 14 chromosomes but their genomes may be 10–15 Mb larger than the *P. falciparum* genome, possibly as a result of differences in the amount of subtelomeric non-coding DNA. Four synteny groups common to all four species were identified, which suggests that gene order has been preserved across species in many cases. Because *P. vivax* is the second most important human malaria and exhibits numerous biological characteristics that differ from *P. falciparum*, it is quite likely that the *P. vivax* genome will be sequenced; an EST gene discovery project has already

been initiated. Comparison of the *P. falciparum* and *P. vivax* genomes should enable the identification of genes responsible for the biological and pathogenicity differences between the two species. In addition, sequence data from murine *Plasmodia* and related parasites such as *Toxoplasma* (Table 1) and *Theileria* [44•] will help to define apicomplexan specific genes.

Conclusions

Tremendous progress towards an understanding of *Plasmodium* biology has been made over the past decade. We can expect the rate of progress to increase in the next decade once the complete genome sequence of *P. falciparum* is determined. This information, coupled with improvements in areas such as informatics, transfection technology, and the development of oligonucleotide [45] and glass slide microarrays [46] for examination of gene expression on a genome-wide scale, will allow investigators to delve into areas of *Plasmodium* biology that are so far unexplored. These discoveries will provide a much more complete picture of malaria parasite biology and facilitate the development of new drugs and vaccines to combat malaria.

Note added in proof

An important new work on *P. falciparum* restriction mapping has just been published [48••].

Acknowledgements

I thank my colleagues at The Institute for Genomic Research (TIGR) and the Naval Medical Research Center (NMRC) for their support. Sequencing of the *P. falciparum* genome at TIGR and the NMRC is supported by the National Institutes of Health, the Burroughs Wellcome Fund, and the Departments of the Navy and Army.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1 World Health Organization: **World malaria situation in 1994: population at risk.** *Wkly Epidemiol Rec* 1997, **72**:269-276.

2. • Sherman IW (Ed): *Malaria Parasite Biology, Pathogenesis, and Protection.* Washington, D.C.: ASM Press; 1998.
This book contains a collection of reviews on most aspects of malaria parasite biology and research.

3. Clayton RA, White O, Fraser CM: **Findings emerging from complete microbial genome sequences.** *Curr Opin Microbiol* 1998, **1**:562-566.

4. Dame JB, Arnot DE, Bourke PF, Chakrabarti D, Christodoulou Z, Coppel RL, Cowman AF, Craig AG, Fischer K, Foster J *et al.*: **Current status of the *Plasmodium falciparum* genome project.** *Mol Biochem Parasitol* 1996, **79**:1-12.

5. •• Su XZ, Wellems TE: ***Plasmodium falciparum*: assignment of microsatellite markers to chromosomes by PFG-PCR.** *Exp Parasitol* 1999, **91**:367-369.
A description of the *P. falciparum* linkage map produced using microsatellite markers covering most of the genome.

6. Reddy GR, Chakrabarti D, Schuster SM, Ferl RJ, Almira EC, Dame JB: **Gene sequence tags from *Plasmodium falciparum* genomic DNA fragments prepared by the 'genease' activity of mung bean nuclease.** *Proc Natl Acad Sci USA* 1993, **90**:9867-9871.

7. Chakrabarti D, Reddy GR, Dame JB, Almira EC, Laipis PJ, Ferl RJ, Yang TP, Rowe TC, Schuster SM: **Analysis of expressed sequence tags from *Plasmodium falciparum.*** *Mol Biochem Parasitol* 1994, **66**:97-104.

8. •• Wellems TE, Su X, Ferdig M, Fidock DA: **Genome projects, genetic analysis, and the changing landscape of malaria research.** *Curr Opin Microbiol* 1999, **2**:415-419.
A review article from one of the leading malaria research groups providing their view of the anticipated impact of recent technological developments, including genome sequencing, on research leading to new drugs and vaccines against malaria.

9. Butler D: **Funding assured for international malaria sequencing project.** *Nature* 1997, **388**:701.

10. Hoffman SL, Bancroft WH, Gottlieb M, James SL, Bond EC, Stephenson JR, Morgan MJ: **Funding for malaria genome sequencing.** *Nature* 1997, **387**:647.

11. Su XZ, Wellems TE: **Toward a high-resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats.** *Genomics* 1996, **33**:430-444.

12. •• Jing J, Aston C, Zhongwu L, Carucci DJ, Gardner MJ, Venter JC, Schwartz DC: **Optical mapping of *Plasmodium falciparum* chromosome 2.** *Genome Res* 1999, **9**:175-181.
A report describing the rapid generation of restriction maps for an entire chromosome by direct visualization of restriction enzyme digested chromosome fragments on glass slides.

13. Aston C, Mishra B, Schwartz DC: **Optical mapping and its potential for large-scale sequencing projects.** *Trends Biotechnol* 1999, **17**:297-302.

14. •• Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C *et al.*: **Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum.*** *Science* 1998, **282**:1126-1132.
This article and the following article by Bowman *et al.* [15••] describe the methods used to sequence the first two *P. falciparum* chromosomes and summarize the major findings.

15. •• Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T *et al.*: **The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum.*** *Nature* 1999, **400**:532-538.
See annotation [14••].

16. Salzberg SL, Pertea M, Delcher A, Gardner MJ, Tettelin H: **Interpolated Markov models for eukaryotic gene finding.** *Genomics* 1999, **59**:24-31.

17. • Newbold CI: **Antigenic variation in *Plasmodium falciparum*: mechanisms and consequences.** *Curr Opin Microbiol* 1999, **2**:420-425.
A concise summary of the progress being made towards understanding the process of antigenic variation in malaria parasites.

18. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M, Taraschi TF, Howard RJ: **Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes.** *Cell* 1995, **82**:77-87.

19. Su Z, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Petersen DS, Ravetch J, Wellems TE: **The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes.** *Cell* 1995, **82**:89-100.

20. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Petersen DS, Pinches R, Newbold CI, Miller LH: **Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes.** *Cell* 1995, **82**:101-110.

21. Borst P, Bitter W, McCulloch R: **Antigenic variation in malaria.** *Cell* 1995, **82**:1-4.

22. Carcy B, Bonnefoy S, Guillotte M, Le SC, Grellier P, Schrevel J, Fandeur T, Mercereau-Puijalon O: **A large multigene family expressed during the erythrocytic schizogony of *Plasmodium falciparum.*** *Mol Biochem Parasitol* 1994, **68**:221-233.

23. •• Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, Lanzer M, Saul A: ***stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens.** *Mol Biochem Parasitol* 1998, **97**:161-176.
A report characterizing novel multigene families encoding proteins involved in antigenic variation.

24. •• Kyes SA, Rowe JA, Kriek N, Newbold CI: **Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum.*** *Proc Natl Acad Sci USA* 1999, **96**:9333-9338.
An article that provides the biological and immunological evidence supporting the classification of the rifins as a novel family of variant surface antigens.

25. Waller RF, Keeling PJ, Donald RGK, Striepen B, Handman E, Lang
•• Unnasch N, Cowman AF, Besra GS, Roos DS, McFadden GI: **Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*.** *Proc Natl Acad Sci USA* 1998, **95**:12352-12357.
Several nuclear-encoded proteins from *Toxoplasma* and *Plasmodium* that are transported into the apicoplast are described and the apicoplast targeting sequences are identified. Also, a new drug target expressed in the apicoplast is described.

26. Wilson RJM, Denny PW, Preiser PR, Rangachari K, Roberts K, Roy A, Whyte A, Strath M, Moore DJ, Moore PW *et al.*: **Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*.** *J Mol Biol* 1996, **261**:155-172.

27. Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJM, Palmer JD, Roos DS: **A plastid of probable green algal origin in apicomplexan parasites.** *Science* 1997, **275**:1485-1489.

28. Roos DS, Crawford MJ, Donald RG, Kissinger JC, Klimczak LJ, Striepen B: **Origin, targeting, and function of the apicomplexan plastid.** *Curr Opin Microbiol* 1999, **2**:426-432.

29. Denny P, Preiser P, Williamson D, Wilson I: **Evidence for a single origin of the 35 kb plastid DNA in apicomplexans.** *Protist* 1998, **149**:51-59.

30. Lang-Unnasch N, Reith ME, Munholland J, Barta JR: **Plastids are widespread and ancient in parasites of the phylum Apicomplexa.** *Int J Parasitol* 1998, **28**:1743-1754.

31. Soldati D: **The apicoplast as a potential therapeutic target in *Toxoplasma* and other apicomplexan parasites.** *Parasitol Today* 1999, **15**:5-7.

32. McFadden GI, Roos DS: **Apicomplexan plastids as drug targets.** *Trends Microbiol* 1999, **7**:328-333.

33. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C,
•• Hintz M, I Tr, Eberl M, Zeidler J, Lichtenthaler HK *et al.*: **Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs.** *Science* 1999, **285**:1573-1576.
This report documents the discovery, using genome sequence information, of a novel drug target in the apicoplast. This is perhaps the most spectacular example to date of the use of genome sequence data to identify new drug targets in *Plasmodium* and related parasites.

34. Ridley RG: **Planting the seeds of new antimalarial drugs.** *Science* 1999, **285**:1502-1503.

35. Roberts F, Roberts CW, Johnson JJ, Kyle DE, Krell T, Coggins JR,
• Coombs GH, Milhous WK, Tzipori S, Ferguson DJ *et al.*: **Evidence for the shikimate pathway in apicomplexan parasites.** *Nature* 1998, **393**:801-805.
The shikimate pathway of chorismate biosynthesis is found in plants, algae, fungi, and bacteria. Chorismate is essential for folate biosynthesis in these organisms and compounds that inhibit enzymes of the shikimate pathway have antimicrobial and herbicidal properties. This paper demonstrates that there is a functional shikimate pathway in *Plasmodium* and related parasites. Because this pathway is not found in mammals, the enzymes of the shikimate pathway may represent new chemotherapeutic targets for antimalarial drugs.

36. Ridley RG: **Planting new targets for antiparasitic drugs.** *Nat Med* 1998, **4**:894-995.

37. Woodrow CJ, Penny JI, Krishna S: **Intraerythrocytic *Plasmodium falciparum* expresses a high affinity facilitative hexose transporter.** *J Biol Chem* 1999, **274**:7272-7277.

38. Miller LH, Hoffman SL: **Research toward vaccines against malaria.** *Nat Med* 1998, **4**:520-524.

39. Good MF, Doolan DL: **Immune effector mechanisms in malaria.** *Curr Opin Immunol* 1999, **11**:412-419.

40. Hoffman SL, Rogers WO, Carucci DJ, Venter JC: **From genomics to**
•• **vaccines: malaria as a model system.** *Nat Med* 1998, **4**:1351-1353.
This article points out that new, high-throughput strategies for the identification and testing of potential vaccine targets must be devised if the full benefits from the enormous amounts of sequence data being generated by the Malaria Genome Project are to be realized. It is proposed that DNA vaccine technology can be used to determine the stage-specificity and subcellular localization of proteins predicted from the genome sequence, and that this information can be used to select antigens for vaccine development.

41. Barry MA, Lai WC, Johnston SA: **Protection against mycoplasma infection infection using expression-library immunization.** *Nature* 1995, **377**:632-635.

42. Golenda CF, Li J, Rosenberg R: **Continuous *in vitro* propagation of the malaria parasite *Plasmodium vivax*.** *Proc Natl Acad Sci USA* 1997, **94**:6786-6791.

43. Carlton JM, Galinski MR, Barnwell JW, Dame JB: **Karyotype and**
• **synteny among the chromosomes of all four species of human malaria parasite.** *Mol Biochem Parasitol* 1999, **101**:23-32.
This article describes the use of pulsed field gel electrophoresis to produce karyotype maps of all four species of malaria parasites that infect humans. All four species appeared to contain 14 chromosomes, but the chromosomes of *P. vivax*, *P. ovale*, and *P. malariae* were found to be larger than those of *P. falciparum*. Four of five synteny groups that are conserved between *P. falciparum* and rodent malarias were conserved in all four human malaria parasites despite the differences in chromosome size.

44. Nene V, Morzaria S, Bishop R: **Organisation and informational**
• **content of the *Theileria parva* genome.** *Mol Biochem Parasitol* 1998, **95**:1-8.
Theileria parva is a cattle parasite that is transmitted by ticks and is related to *Plasmodium* and *Toxoplasma*. *T. parva* infects the lymphocytes of the host and causes a lymphoproliferative disorder called East Coast Fever. This article reviews current knowledge of the *T. parva* genome, which is one-third the size of the *P. falciparum* genome. Sequencing of *T. parva* genome would assist the characterization of the malaria parasite genome and may also shed light on the mechnisms of *T. parva*-induced host-cell transformation.

45. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21**:20-24.

46. Debouck C, Goodfellow PN: **DNA microarrays in drug discovery and development.** *Nat Genet* 1999, **21**:48-50.

47. Ajioka JW, Boothroyd JC, Brunk BP, Hehl A, Hillier L, Manger ID, Marra M, Overton GC, Roos DS, Wan KL *et al.*: **Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa.** *Genome Res* 1999, **8**:18-28.

48. Lai, Z, Jing J, Aston C, Clarke V, Apodaca J, Dimalanta ET, Carucci DJ,
•• Gardner MJ, Mishra B, Anantharaman TS *et al.*: **A shotgun optical map of the entire *Plasmodium falciparum* genome.** *Nat Genet* 1999, **23**:309-313.
This is an extension of the work reported by Jing *et al.* [12••], where optical restriction mapping was used to rapidly prepare a restriction map of *P. falciparum* chromosome 2. In this paper, an optical restriction map of the complete *P. falciparum* genome was constructed. Optical maps of *P. falciparum* chromosomes have proven very useful for gap closure and sequence verification. Optical restriction maps may be very useful in the sequencing of *Plasmodium* genomes for which other physical or genetic maps are not available.